DOCUMENT RESUME

ED 285 919 TM 870 644

AUTHOR	Brown, George H.; Faupel, Elizabeth M.
TITLE	The National Assessment of Educational Progress and the Longitudinal Studies Program: Together or Apart? Report of a Planning Conference (Washington, DC, December 11, 1986).
INSTITUTION	Center for Education Statistics (OERI/ED), Washington, DC.
REPORT NO	CS-87-446
PUB DATE	87
NOTE	141p.; For reports from related conferences, see HE 020 862 and PS 016 941.
AVAILABLE FROM	Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402.
PUB TYPE	Collected Works - Conference Proceedings (021) -- Viewpoints (120)
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	Conferences; *Cross Sectional Studies; Databases; Data Collection; *Educational Assessment; Educational Trends; Elementary Secondary Education; Federal Programs; *Longitudinal Studies; Mergers; *National Surveys; Outcomes of Education; *Policy Formation; Research Design; Research Utilization; Sampling; Testing Programs
IDENTIFIERS	Center for Education Statistics; Longitudinal Studies Program; *National Assessment of Educational Progress; *National Education Longitudinal Study 1988

ABSTRACT

The National Assessment of Educational Progress (NAEP) and the Longitudinal Studies Program (LSP) are major survey projects on educational outcomes performed by the Center for Education Statistics. NAEP is a continuing cross-sectional survey of young Americans' skills, knowledge, and attitudes. The LSP studies follow a sample of students as they progress through school into work and family life. This document reports on a planning conference to develop recommendations for the Center for Education Statistics on merging NAEP and the National Education Longitudinal Study (NELS) of 1988. Both the technical problems and complex ramifications of the merger were addressed. Summaries and full texts of the five conference papers which had been commissioned by expert panelists are presented: (1) "More Bang for the Buck: An Integrated Data Collection Strategy" (Alan L. Ginsburg et al.); (2) "Shooting at a Moving Target: Merging the National Assessment of Educational Progress and the Longitudinal Studies Program--A State Perspective" (Joan Boykoff Baron and Pascal D. Forgione); (3) "How to Optimize and Articulate a Longitudinal and a Cross Sectional Research Program" (Calvin C. Jones); (4) "Instrument Design for a Combined NAEP and NELS" (R. Darrell Bock); and (5) "Sampling Problems in Merging a Cross-Sectional and a Longitudinal Program" (Bruce D. Spencer). Summary remarks by David Sweet and Emerson Elliott of the Center are presented. Papers written after the conference by senior officials from the Center include discussions of implications by David A. Sweet, Gary W. Phillips, and C. Dennis Carroll. (BS)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

The National Assessment of Educational Progress and the Longitudinal Studies Program: Together or Apart?

Report of a Planning Conference

December 11, 1986

Prepared by
George H. Brown
with
Elizabeth M. Faupel

U.S. Department of Education
William J. Bennett
Secretary

Office of Educational Research and Improvement
Chester E. Finn, Jr.
Assistant Secretary

Center for Education Statistics
Emerson J. Elliott
Director

Information Services
Edwin S. Darrell
Director

Center for Education Statistics

"The purpose of the Center shall be to collect and disseminate statistics and other data related to education in the United States and in other nations. The Center shall . . . collect, collate, and from time to time, report full and complete statistics on the conditions of education in the United States; conduct and publish reports on specialized analyses of the meaning and significance of such statistics; . . . and review and report on education activities in foreign countries,"–Section 406(b) of the General Education Provisions Act, as amended (20 U.S.C. 1221e-1).

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402

CONTENTS

Chapter	Page
1. Introduction	1
2. Paper Summaries	5
3. Summary of Discussions	21
4. Implications for the Center	25
A. Comments from the NELS Perspective, by C. Dennis Carroll	27
B. Comments from the NAEP Perspective, by Gary W. Phillips	35
C. Comments on the Conference, by David A. Sweet	39
Appendixes:	
A. Commissioned Papers	
(1) Getting More Bang for the Buck: Increasing the Utility of Education Data Bases, by Alan L. Ginsburg, Valena Plisko, Nora Guhl, and David Myers	43
(2) Shooting at a Moving Target: Merging the National Assessment of Educational Progress and the Longitudinal Studies Program--A State Perspective, by Joan Boykoff Baron and Pascal D. Forgione	63
(3) How to Optimize and Articulate a Longitudinal and a Cross Sectional Program, by Calvin C. Jones	73
(4) Instrument Design for a Combined NAEP and NELS, by R. Darrell Bock	99
(5) Sampling Problems in Merging a Cross Sectional and a Longitudinal Program, by Bruce D. Spencer	117
B. List of Conference Participants	141
C. Biographical Sketches of Speakers	145

Chapter 1

Introduction

The National Assessment of Educational Progress (NAEP) and the Longitudinal Studies Program (LSP) are two large and important projects done by the Center for Education Statistics in the Office of Educational Research and Improvement, U.S. Department of Education. Both studies concern education outcomes in the broad sense but their objectives are different.

NAEP is a congressionally mandated continuing survey of the skills, knowledge, and attitudes of young Americans. Data collection began in 1969. During the last 18 years NAEP assessed over 1 million 9-, 13-, and 17-year-olds and young adults and reported their performance and performance trends.

The areas assessed were writing, reading, mathematics, science, literature, art, music, social studies, computer competence, citizenship, and career and occupational development. The objectives for each assessment are based on the consensus of citizens across the Nation. Test items (or "exercises" as they are called in NAEP) are developed by subject matter experts. The skills and content areas vary from one assessment to another depending on the priorities of policy officials and information needed by curriculum planners and authors of tests and text books.

NAEP has always been a cross sectional survey and uses a technique known as matrix sampling. There are no scores for individual students. It is not possible in the NAEP program to follow a student over time to examine the effects of school, family, or other factors on educational achievement and vocational outcomes.

The Center's Longitudinal Studies Program began with the National Longitudinal Study of the High School Class of 1972 (NLS-72) and continued with High School and Beyond (HS&B) in 1980 with surveys of high school sophomores and seniors. HS&B has had three follow-up surveys at two-year intervals. In 1988, the Center's third major longitudinal study, National Education Longitudinal Study of 1988 (NELS:88), will begin with a base-year survey of eighth grade students.

In these longitudinal studies, students are selected to constitute a representative sample of their grade levels and data are obtained on each student. Students are followed as they progress through school and make their transition to work and family life. Education attainment is measured by years of school completed, courses taken and grades earned. Because of time constraints, achievement testing covers fewer content areas than are covered by NAEP.

Over the last 15 years, people in and outside the Department of Education raised the question whether substantial benefits might result from combining these two large projects which cover overlapping age groups. Informal discussions have taken place over the years. But this conference is an effort to examine the question in depth by people who are knowledgeable about the technical problems and complex ramifications that would result from a merger of NAEP and LSP.

The conference addressed these questions: (1) Is a merger of NAEP and LSP desirable?, and (2) If so, what problems would result from a merger and how might they be resolved?

It is a propitious time for this conference for two reasons: (1) the structure and character of NAEP were examined by the Alexander-James Study Group, to recommend any changes. (2) The longitudinal study, NELS:88, is not fully underway--the base-year data collection will take place in 1988. If NAEP and LSP merge, they should do so before launched on their separate courses.

Selection of Panel Members

A distinguished panel was selected from a list of experts recommended to the Department by individuals who themselves had excellent credentials in longitudinal and cross sectional research methodology, assessment methods, psychometrics, and sampling theory.

The panelists' charge was to address the following issues:

> Are there ways of enhancing the value of NAEP, LSP, or both by bringing the two closer? Should NELS:88, a longitudinal study, incorporate full-scale assessments such as are carried out by NAEP? Should a regular NAEP sample be converted to a panel for longitudinal assessment? If so, how could it be done given the Balanced Incomplete Block Spiralling (BIBS) sampling procedure? How should NAEP and LSP relate to the new elementary-secondary data system?

The panelists were commissioned to prepare and present papers about an aspect of the conference topic. The conference schedule gave ample time for discussion of each presentation. Summarizing and synthesizing the presentations was done by senior officials of the Center for Education Statistics at the conference and in the commentary papers which compose chapter 4 of this report.

The overall objective of the conference, namely, to develop recommendations for the Center for Education Statistics on merging NAEP and NELS:88, was emphasized with each speaker.

Plan of This Report

Chapter 2 has summaries of the five commissioned papers--to give the reader a sense of the conference. These are the five papers and their authors:

* Getting More Bang for the Buck: Increasing the Utility of Education Data Bases, by Alan L. Ginsburg and Valena Plisko, Office of Planning, Budget, and Evaluation, U.S. Department of Education, and Nora Guhl and David Myers, Decision Resources, Inc.

* Shooting at a Moving Target: Merging the National Assessment of Educational Progress and the Longitudinal Studies Program--A State Perspective, by Joan Boykoff Baron and Pascal D. Forgione, Connecticut State Department of Education.

* How to Optimize and Articulate a Longitudinal and a Cross Sectional Program, by Calvin C. Jones, National Opinion Research Center.

* Instrument Design for a Combined NAEP and NELS, by R. Darrell Bock, The University of Chicago.

* Sampling Problems in Merging a Cross-sectional and a Longitudinal Program, by Bruce D. Spencer, Northwestern University.

Chapter 3, Summary of Discussions, comprises three parts: (1) condensed statements of the principal viewpoints from the discussions; (2) summary remarks by David Sweet, Director of the Education Outcomes Division at the Center; and (3) summary remarks by Emerson Elliott, Director of the Center.

Chapter 4, "Implications for the Center," comprises three papers prepared after the conference by Center personnel:

* Comments from the NELS Perspective, by C. Dennis Carroll, Chief, Longitudinal Studies Branch, Education Outcomes Division.

* Comments from the NAEP Perspective, by Gary W. Phillips, Chief, Cross Sectional and Special Studies Branch, Education Outcomes Division.

* Comments on the Conference, by David A. Sweet, Director, Education Outcomes Division.

The appendix has the five papers provided by the authors. In some instances, a paper was slightly revised by the author after presentation. The appendix also has a list of the conference participants and biographical sketches of panel members.

Chapter 2

Paper Summaries

This chapter presents summaries of the five papers commissioned for the conference. Each summary was prepared by the author.

More Bang for the Buck: An Integrated Data Collection Strategy

Alan Ginsburg, Nora Guhl, David Myers, and Val Plisko

Federal data collection activities in education are undergoing extensive review. One of the key questions facing policymakers is the design of the blueprint for collection of data involving comparisons of students' educational performance over time and the related variables that affect this performance.

Repeated measures of student educational performance over time, either longitudinal (collecting data from the same student) or cross-sectional (collecting data from different students), serve several important functions for educational policymaking: (1) over-time data are important for monitoring trends in student educational performance and (2) over-time data are important for assessing the short- and long-term effects of educational programs and background attributes on student performance.

Although repeated cross-sectional and longitudinal data yield important policy-relevant information about educational outcomes and processes, both types of data are expensive to collect from nationally representative samples. With funding for data collection by the Department of Education likely to be limited, ensuring efficiency in data collection and use assumes great importance. This paper considers ways to expand the use of extant data bases as well as ways to improve future data collection efforts. Our recommendations concern three main areas: improving data utilization; enhancing the quality of information; and enhancing the capacity of national surveys to identify causal relationships.

One of the ways in which data use can be improved is to encourage greater application of data bases produced outside the Department of Education to educational concerns. More than $100 million has been invested by other Federal agencies in producing data bases that can be used to analyze educational issues. Thus far, these data bases have not been used extensively by educational researchers. The Department of Education can support the use of "outside" data bases by funding studies that extend analysis beyond Department of Education data or by encouraging participation in education conferences by experts who are knowledgeable about "outside" data bases.

Another way to improve data use is to prepare user tapes with clear and detailed documentation for potential users. Where these have been prepared, such as for the major longitudinal surveys from the Education and Labor Departments, literally hundreds of studies have been carried out. This is in stark contrast to the National Assessment of Educational Progress and the Sustaining Effects Study of Title I which, despite their $50 million+ cost, have gone virtually unused by researchers other than the federally funded contractors that collected the data.

Small, competitively awarded Federal grants also could increase the yield derived from multimillion-dollar investments.

In addition to expanding the use of extant data, it is important for the Department of Education to enhance the quality of information it collects by exercising greater quality control over the reliability and validity of measures. Where feasible, the Department of Education should periodically check data it collects for consistency with other sources of similar data. Also, the Department of Education should devote greater attention to collection of information from the appropriate respondent. In the case of student's family background, this may necessitate the added expense of a household survey.

It is also important to obtain more information on the schooling or home processes that give rise to the observed outcomes. Specialized surveys such as NAEP, which historically focused primarily on obtaining outcome information, could be made more useful if supplemented with questions about the content of the coursework students take. The International Association for the Evaluation of Educational Achievement's math survey, SIMS, which explored the "opportunity to learn" as a predictor of student achievement, serves as an important model of the ways in which information about content can be obtained.

The Department of Education also needs to direct its efforts towards enhancing the capacity of national surveys to identify causal relationships. Cross-sectional studies such as NAEP should be supplemented with smaller subsamples of students who are followed longitudinally. The subsample could represent all students or follow a particular population of interest, such as students from a particular grade, or a group of programmic interest.

Retrospective information should be collected on respondents when reliability is satisfactory. Retrospective data cost relatively little time and money to collect compared with true longitudinal data. Administrative records, such as transcripts, afford a relatively accurate means for obtaining a student's academic history. Respondent recall may yield adequate responses for some kinds of information of high policy interest, such as work or marriage history. At a minimum, the Department of Education should fund studies of the accuracy of different types of retrospective information.

In closing, although our discussion has centered on the improvement of large-scale national data collections, we have recommended that small-scale studies may be appropriate supplements or alternatives to large-scale studies in a number of cases. Our overriding concern is that we have research tools that enable us to address the issues that face educational policymakers. Given the paramount importance of learning about the effects of educational programs as well as the concern about cost, educational policymakers may have to consider greater use of small-scale, detailed case studies.

Summary

Shooting at a Moving Target: Merging the National Assessment of Educational Progress and the Longitudinal Studies Program--A State Perspective

Joan Boykoff Baron and Pascal D. Forgione, Jr.
Connecticut State Department of Education
and
Marc Moss
Harvard University

It is the view of members of the Connecticut State Department that a merger of NAEP and NELS is advisable if it can be achieved without sacrificing the Nation's ability to make longitudinal comparisons with past data collected in these two programs. This conclusion is based on separate analysses of the advantages and disadvantages of such a merger to stakeholders at the Federal, State and local levels and whether any of the original purposes of the separate programs would be sacrificed in a merged program. These analyses indicate that the primary advantage of a merger would be a reduction in the amount of testing and disruption in our nation's schools. The anticipated testing configuration resulting from the combined effects of NELS, NAEP, the proposed CCSSO assessment program, and annual state testing programs is a large administrative burden in certain schools at certain grades. The significance of the title is that this paper was updated several times in line with important decisions with major testing ramifications that were made in rapid succession of one another. The paper was last updated following the recommendations of the Study Group on the National Assessment of Student Achievement chaired by Lamar Alexander and further represents an attempt to integrate those reflections into the present framework for a merger of NELS and NAEP.

The paper presents two possible models for accomplishing a merger which differ on the extent to which they provide for cross-sectional or cohort analyses. The ideal merger would study both the performance of a representative cross section of students at designated grade levels at regularly scheduled intervals as well as the performance of the same cohort of students over time. Such a merger would enable the isolation of the effects of any confound that might result from a "Hawthorne effect" emanating from an in-depth longitudinal analysis on the same students and schools.

The authors also suggest several additional analyses that might be conducted if such a merger were achieved. This assumes that one large study would be less costly than two smaller studies and funds would be available to expand into important new areas in the merged study. One important area would be the

inclusion of classroom observations which would enable researchers to supplement achievement data and self-report data with classroom process data. This would permit a better understanding of the relationship between instruction and performance. A second area would be to develop and test causal models related to both the predictors and results of different levels of achievement. A third area is that of item response theory which has been used successfully in several large scale assessments but whose potential has not yet been tapped at the national level.

Though desirable from a logistical perspective, the realization of such a merger ultimately depends on its technical feasibility. It is critically important that a merger not interfere with our nation's ability to chart its educational progress over time. Therefore, once a psychometrically sound technical plan is developed, it will be important for representatives from the states to review and comment upon it.

Summary

How to Optimize and Articulate a Longitudinal and a Cross Sectional Research Program

Calvin C. Jones
National Opinion Research Center

Both the National Assessment of Educational Progress (NAEP) and the National Education Longitudinal Studies (NELS) programs trace their origins to the late 1960's. At that time the (National) Center for Education Statistics (CES), the primary data gathering and reporting unit within the U.S. Department of Education, was being pressed by both the executive and legislative branches to collect data on a national and regional basis that would improve understanding of the dynamic properties of the U.S. elementary and secondary school system.

At its inception, each of these two programs had its own political and research constituencies marshalling resources and forging the design and methods to address a particular set of issues. The unique scientific, methodological, and operational problems facing each of the new enterprises were numerous and formidable. Many talented, dedicated individuals in both the private and public sectors demonstrated enormous creativity in developing solutions appropriate for the times and for the distinctive needs of each study. The result, nearly two decades later, is a pair of highly specialized research programs, each demanding substantial dollar resources on a recurring basis, with relatively little naturally occurring articulation between them, and with a strong tendency within each to expand and build upon its own roots, and to resist gratuitous or fashionable change.

NELS and NAEP. The fundamental purpose of the longitudinal NELS surveys has been the study of stability and change in the educational activities of American youth, including the extent to which students' background characteristics and other exogenous variables affect educational achievement, and the subsequent impact of educational activities and attainment on such other outcomes as labor force participation and economic well-being, family formation, military service, and citizenship. The program consists of three studies: The National Longitudinal Study of the High School Class of 1972 (NLS-72), High School and Beyond (HS&B), and the National Education Longitudinal Study of 1988 (NELS:88).

Central to the cross sectional NAEP's research objectives and design is the repeated measurement, at short, regular intervals, of the level of developed ability of elementary and secondary school students in basic subjects such as reading, mathematics, writing, science, and social studies. By testing

three age (and/or grade) cohorts (most commonly 9-, 13-, and 17-year-olds) in each assessment, NAEP permits analyses of the differences in developed abilities among the three groups. By assessing the same subject areas at intervals, data from the time series of NAEP surveys permit analyses of trends in what each age cohort has learned or can do, and analyses of changes over time in the differences in performance or capabilities among the age groups measured, and subgroups within each age cohort.

Despite their differences in purpose, method, and allocation, both studies spend a great deal of their resources on similar activities that, if coordinated and integrated, might release funds for improvements in methods and measurement tools, as well as in the quantity and quality of the data they collect. Of the infinite number of models that might be examined for articulating NELS and NAEP, it is perhaps most instructive, at this point, to consider two extreme types:

> Making comparatively minor changes in the design of each program, while retaining the separate identities and unique characteristics of each.
>
> Completely merging every feasible operation of the two programs into a unified study that attempts to accommodate the full set of objectives pursued by both using a single sample for each cohort studies.

Relatively low levels of coordination, such as establishing a common population definition for the student cohorts studied, will not in themselves result in any economies or have profound effects upon the overall research value of either study. However, it is difficult to imagine the path to much higher and more cost effective levels of integration including the extreme of collecting assessment and longitudinal data from the same students until these lower level issues are resolved in favor of consistency between the programs.

Opportunities for Integration. Major areas in which the two programs might move toward integration are (1) population definitions and sample design, for students, school dropouts, and districts and schools, (2) both general data collection strategies, such as access to schools, and individual data collection and testing and assessment strategies. The opportunities for integration in general data collection strategy, and the possible savings from such an approach, will illustrate the benefits of even limited integration.

Both NAEP and NELS use multi-stage probability sample designs. That is, students are not sampled directly but are randomly chosen from the rosters of enrolled students in nationally representative samples of schools. For the sake of sample design efficiency, the number of clusters (schools) selected in the two programs is large--typically between 700 and 900 schools for each NAEP cohort and about 1,000 schools for NELS.

This means that both studies must spend a great deal of time and effort on the process of gaining the voluntary cooperation and assistance of state education agencies and large numbers of districts and schools in which their student samples are enrolled. Once cooperation is obtained, teams of survey administrators must travel to the school sites to collect questionnaire and test data from students in group sessions. Cost data from NAEP and NELS indicate that approximately 60 percent of the total student and data collection resources for each program are consumed by the process of setting up school-based administrations, with only about 40 percent dedicated specifically to collecting and processing student responses.

Although there is considerable overlap between the programs at the district level, the number of overlapping schools is quite small. In a year when both NAEP and NELS are in the field, there may be as many as 2,700 separate schools selected for the two, consuming between $2 and $3 million for access and setup before the first student is surveyed. If the organization of data collection for the two programs were combined--even if different types of data were collected for each--access costs could be substantially reduced, freeing resources for more pressing research problems.

Fully Merging NAEP and NELS. A complete merger of NAEP and NELS will require, to begin, both consistent grade-based population definitions and an assessment and test design acceptable for the two programs. Having these two elements in place would permit the initiation of a fully merged longitudinal assessment design in the spring of any even-numbered year in the 1990's.

Base year studies for three parallel longitudinal studies of 4th, 8th, and 12th grade cohorts could be conducted in approximately 1,000 public and private schools for each cohort. The total of 3,000 schools for a merged sample is only slightly larger than the expected number of about 2,700 schools to be sampled by the two programs in 1988. Furthermore, by 1992 the two separate programs may be operating in over 3,500 schools. In addition, this approach could realize the economies possible in combined efforts to secure school cooperation, discussed above as part of the more limited approach to integration.

An integrated approach to assessment and testing might be achieved through the development of an entirely new test that would abandon past approaches to assessment and longitudinal measurement of growth and attempt to serve both purposes with a new framework. Use of the "duplex test design" developed by Bock, Mislevy, and their colleagues would require a break with the historical continuity of past assessments and longitudinal studies, but it would offer a good deal for a combined approach to the two programs. It is clearly more feasible than some other alternatives; it shares some of the item sampling techniques of

the traditional NAEP approach; and it offers a means for evaluating the progress of schools as well as students toward meeting established curriculum objectives. Running perfectly counter to NAEP's earliest orientation, the duplex design would begin to provide assessment data useful for improving schools that so many educators, administrators, and researchers have been demanding.

Furthermore, the duplex design offers flexibility in testing procedure. Configured as a two-stage test, the duplex design can be and has been administered using printed booklets and standard group testing procedures. However, as an adaptive design, its efficiency could be substantially increased through computerization. If the properties and resulting scores from the duplex test design are preferred to traditional NAEP and NELS test designs, use of this testing approach would be compatible with computerized data collection at survey and testing centers should such a procedure prove feasible.

Even if the elements of a fully merged design could be settled for the base year, however, many challenging complications would arise for the next assessment or longitudinal followup. Various possible approaches present distinct disadvantages, including threats to the very purposes of the design change, increased costs, added problems in tracking respondents, and unacceptable increases in demands on respondent recall. There are alternatives without these costs, but they require significant departures from current practice.

Conclusion. By combining the design strengths of NELS with a testing program suitable for assessment reporting, we could move much closer to the goal of understanding what works in educational reform. A merged design should also preserve other features of the most recently developed longitudinal study, namely, the collection of data from parents, teachers, and school principals of all sampled students, and should include the collection and processing of report cards, transcripts, and other student records. The data structures provided by this design would come very close to satisfying the long-range goals and requirements of a truly integrated system for studying elementary and secondary education. However, these ambitions may never be fulfilled unless the first, small steps toward articulating the longitudinal and assessment programs are taken in the very near future.

Summary

Instrument Design for a Combined NAEP and NELS

R. Darrell Bock
University of Chicago

NAEP and NELS have traditionally collected much different types of data. Evolved out of the school accountability movement, NAEP has used matrix-sampling methods in order to assess the progress of the nation's schools in meeting detailed curricular objectives at the 4th, 7th, and 11th grade level. NELS, using conventional achievement tests, collects longitudinal data for the developmental study of learning and attainment of individual students. It would be possible to combine these data gathering efforts only if an instrument were designed that provided for both assessment of curricular objectives at the school level and measurement of individual achievement at the student level. Such an instrument, called a "duplex" design, has recently been proposed and is being tested in Illinois and California by the OERI Center for Student Testing, Evaluation and Standards. Although there are advantages to NAEP and NELS employing a common instrument of this type, the conflicting requirements of a cross-sectional study (NAEP) and a longitudinal study (NELS) make it difficult to base the two efforts on one body of data collected from the same respondents. For the present, integration of the two studies probably could not proceed beyond the level of cooperation in those phases of instrument design that would allow the results of the two studies to be related conceptually.

Sampling Problems in Merging a Cross-Sectional and a Longitudinal Program

Bruce D. Spencer
Associate Professor of Statistics, Education and Social Policy, and Urban Affairs and Policy Research
Northwestern University
and
Director, Methodology Research Center, NORC

Four kinds of sampling problems in merging a longitudinal and a cross-sectional program are discussed:

1. Maintaining representative cross-sections;
2. Maintaining acceptable levels of nonresponse;
3. Making NAEP more longitudinal; and
4. Embedding a NELS survey in NAEP.

1. Maintaining Representative Cross Sections

The first general problem in merging a longitudinal and a cross-sectional program is to maintain a representative cross-sectional sample at each time point. (A representative sample is a probability sample from the target population with known selection probabilities for the units in the sample.)

Part of this problem is easily handled. Consider, for example, NELS:88. NELS:88 will select a representative sample of 8th grade students and 8th grade schools (i.e., schools containing 8th grade students) in the 1987-88 school year. These students will be followed up in 1990, when most, but not all, are in 10th grade. To obtain a representative sample of 10th graders in 1990 we will need to do two things. First, we will need to exclude those members of the base-year sample who are not in 10th grade in 1990. Second, we will pull an additional sample of 1990 10th grade students who were not eligible for selection in 1988--these include, for example, immigrants, certain 1988 9th grade students who repeated a grade and 1988 8th grade students whose schools were not eligible for a base-year selection but who are part of the target population of 1990 10th graders. The result will be a probability sample of 1990 10th graders.

There are also problems inherent with maintaining a longitudinal sample of younger students over time. As a sample of student graduates from elementary schools to middle schools or junior high schools to high schools, the number of school buildings in which they attend school will tend to increase. Thus, if 28,000 students in 1,000 schools composed a 1988 8th grade sample and if all of the sampled students were resurveyed in school in 1990 it might be necessary to visit more than, say, 2,000 or 3,000 schools. That would be unduly expensive, and subsampling strategies will need to be explored to control field

costs and maintain statistical efficiency. Note, however, that this problem is inherent to longitudinal surveys of students whether or not the longitudinal survey is combined with cross-sectional surveys. Calvin Jones's proposal for the creation of testing centers might ameliorate this problem (Jones, 1986).

2. Maintaining Acceptable Levels of Nonresponse

A second potential sampling problem with merging a longitudinal and a cross-sectional program is the possibility that nonresponse will be greater if a cross-sectional survey is embedded in a longitudinal survey than if repeated independent cross-sectional surveys are conducted. This problem can arise because longitudinal surveys run the risk of cumulative nonresponse by students, leading to increasing levels of nonresponse with each successive wave. In fact, however, each follow-up in the National Longitudinal Survey of the class of 1972 (NLS-72) and High School and Beyond had lower (weighted) nonresponse rates than did the base-year surveys, showing that increases in the number of base-year nonrespondents who participated in later waves can outweigh increases in attrition. On the other hand, each successive follow-up after the first one suffered larger nonresponse than the preceding. The conclusion to be drawn is that nonresponse rates in longitudinal surveys do not have to be higher than nonresponse rates for cross-sectional surveys of students in schools.

3. Making NAEP More Longitudinal

NAEP exemplifies a repeated cross-sectional design. As a primary use of NAEP data is to assess change, both across cohorts and within cohorts, analyses of change would benefit enormously from a more longitudinal design. NAEP could be made more longitudinal in two ways, by keeping schools and school systems in the sample for multiple times or by testing the same students at multiple times. The primary advantage of longitudinal testing of students would be the ability to correlate student growth with school practices and other factors. Such analyses, however, are better performed with NELS surveys, which include more detail on school and home environments. The primary advantages of longitudinal testing in schools and school systems are more precise estimates of change in performance, at the school level, at the district level, at the state level, and nationally.

4. Embedding a NELS Survey in NAEP

A variety of ways exist to merge a NELS survey into the NAEP design. The primary advantage in merging the two samples in this way will stem from the additional achievement data at the school level that NAEP can provide if a design such as Darrell Bock's "duplex design" is used. The duplex design (Bock, 1986) can produce accurate estimates of student ability at the school level

if at least 30 students per school are tested. The increased accuracy may be of some value for studying instructional effects; although one should note that NELS studies are poorly suited for assessing the effectiveness of teaching. As students receive instruction from one teacher for a period of a year, a more powerful method is to test the students prior to the instructional period and at the end of this period (e.g., spring-spring testing, or fall-spring testing) and to correlate the achievement gains with teaching method. Such a design is used in the International Educational Assessment (IEA) surveys. Of course, randomized assignment of teaching method would lead to yet more powerful studies of instruction, but even with such an experimental approach, moving towards the IEA design would be valuable. Thus, estimates of school-wide ability can be provided with high accuracy, which can be of value for the study of school effects.

Various mergers of NAEP and NELS are possible. The simplest mergers involve merging the base-year of NELS with a NAEP sample at the same grade. It would be rather straightforward to merge a NELS 10th grade (base-year) sample with a NAEP 12th grade sample two years later, or to merge a NAEP 4th grade sample with a NELS 6th grade sample two years later. Other mergers appear problematic.

Critical questions that need to be addressed in designing any merger are, What analyses would be possible with a merged program that are not possible without a merger? What would we learn from the data? The kinds of mergers considered here do not promise many analytical benefits unless the studies are strengthened in other ways as well. For example, testing students twice within a 12-month period would allow measurement of growth, which could be statistically associated with teaching methods and other school processes. If such analyses were possible, then the improved testing data afforded by NAEP would have real value for NELS. However, unless we can specify the analytic gains we need from the merger, we might end up with an inferior design.

Chapter 3

Summary of Discussions

This chapter presents in brief, the principal viewpoints expressed in discussions during this conference.

* Serious doubts were expresed that States with an active assessment program or that piggy-back on NAEP or NELS, would be willing to give up much of their own test content to conform to the requirements of a national merged system.

* We need to build political structures that would make it possible to negotiate test content.

* Repeated testing of the same subjects, as is done in longitudinal studies, carries risks including the possibility that schools will start "teaching to the test." This complicates causal analysis. An opposing view was that "teaching to the test" may be a good thing provided schools are teaching and testing the right things.

* Before getting into all the complexities of how to carry out a merger of NAEP and NELS, we should carefully consider and decide whether a merger is desirable.

* Serious doubts were expressed by some participants whether any financial savings would result from a merger. One person said that if a merger were decided on, and it appeared that some savings were likely, then the surplus funds should be used for enhancing the merged program.

* There are serious risks in trying to merge two or three studies that are operational and appear to be functioning well. We could wind up with a huge merged system that does nothing as well as the original studies.

* Perhaps another conference is needed to discuss the advantages and economies of consolidating some of the expensive Government sponsored surveys of disadvantaged groups. Some $10 or $12 million a year are spent on these.

* An unaddressed issue in this conference is the question of the State information needs that might be met by a merged program. We are operating in an information vacuum about use. Perhaps some papers should be commissioned to investigate and clarify this matter.

Summary Remarks by David Sweet

There has been a synergistic relationship between NAEP and NELS for some time, even without a direct link or merger. There has been much sharing of information in methodology, in the importance of gathering information on special populations, and on how to couch issues.

NELS has benefitted from some of the NAEP achievement items. NAEP has recently benefitted from some of the analyses NELS has done of the high school transcripts. There have been many exchanges of that sort. The two staffs talk with each other. The two contractors are meeting, at least here at this conference. This conference may be one of those rare occasions where, even if we are asking the wrong questions, we may be getting very good and useful answers. It may be the wrong question to ask whether or not the two studies should be merged. But the good message that we are hearing is that the two studies can learn from one another and coordination of the two studies is of growing importance.

I think also that many good ideas have come out of this conference that are useful to both projects. I have in mind such things as Bock's duplex design, some of the complications of coordinating the samples and data collection operations, consideration of the burden on the schools, the coordination difficulties of having samples with different purposes at different grade levels. Clearly, both studies will continue to profit from knowing what's going on in the other.

Summary Remarks by Emerson Elliott

It seems to me that we didn't resolve the issue here. It was pretty clear before we started that we wouldn't but it's also clear from this conversation that more people need to get in on this debate because it is a very significant one. What I have been jotting down are themes that seemed to me to come out of this and it's not a straight cut yet. Let me just spit them out because I'd like to encourage you to argue with me if I haven't heard them correctly.

There are different purposes that might be served or that could be served by a merger that have been enunciated at various times. One of the things I would like to have, in addition to the transcript, is a clear description of what they are, with pros and cons about how a merger might address them. There are at least the following: (These are in no sort of order.)

(1) A national report card.
(2) A report on program effects.
(3) A report on school effects.
(4) Reduction in data burden.
(5) Costs.
(6) Correlations between achievement and school characteristics.
(7) Growth or change or gain.

Another thing I heard here, but in a different dimension, is that the issue is not whether to merge or not to merge. In fact, the issue runs all the way from, "Don't do anything about merging" to a complete merger. Between these two extremes are the possibility of tagging short longitudinal studies on to NAEP in a variety of ways, or doing other things short of a full, complete, permanent merger of studies. Each of these options would presumably have its own pros and cons.

It seems to me that these are some of the dimensions that should be highlighted and made more crisp so that other people can join the debate and express their views on these same questions.

Chapter 4

Implications for the Center

This chapter comprises commentaries by three Center officials on the desirability and feasibility of merging NAEP and NELS.

Part A. Comments from the NELS Perspective, by C. Dennis Carroll, Chief, Longitudinal Studies Branch.

Part B. Comments from the NAEP Perspective, by Gary W. Phillips, Chief, Cross Sectional and Special Studies Branch.

Part C. Comments on the Conference, by David A. Sweet, Director, Education Outcomes Division.

Part A

Comments from the NELS Perspective

C. Dennis Carroll, Chief
Longitudinal Studies Branch

The National Assessment of Educational Progress (NAEP) and the National Education Longitudinal Studies (NELS) are two large, useful projects in the Center for Education Statistics. In the early 1970's, NAEP and NELS were distinct with NAEP focused on elementary and secondary students' achievements and NELS (NLS-72) focused on college access and choice. NAEP continued its focus on elementary and secondary students' achievement. However, the second longitudinal study, High School and Beyond, included a cohort of high school sophomores in 1980 and the distinction between NAEP and NELS was slightly eroded. The third longitudinal study, the National Education Longitudinal Study of 1988 (NELS:88), focuses on a cohort of eighth graders in 1988. Hence, the longitudinal program focus has changed from postsecondary education to secondary. As such, it is reasonable to consider if and how NAEP and NELS can be merged, integrated, or structured to provide better data at lower cost with less burden.

To some extent, NAEP and NELS have different constituents. NAEP's cross-sectional snapshots of student achievements have fostered notions of a national report card. Trend comparisons among NAEP's have also allowed some measures of growth (e.g., comparisons of fourth graders in 1972 with eighth graders in 1976). Both of these uses of NAEP data are informative and attractive to elementary/secondary school systems. Just as parents desire good marks on their child's report card (and compare the marks of younger children with older siblings), the NAEP constituents are eager to learn how well typical elementary/secondary students read and compute. Recently, NAEP data has supported the identification relationships of school and student characteristics to student achievements. However, because NAEP is cross-sectional, it is difficult to establish why these relationships exist.

NELS constituents have focused on two unique attributes of longitudinal studies. First, NELS provides information on timing--which allows precursor events to be related to outcomes. Causal orderings may be established and behaviors may be modeled. Hence, NELS speaks to why relationships exist. Second, NELS has focused on change. Recently, cognitive growth has been one of the changes NELS has addressed. However, the majority of the changes NELS has encompassed have been transitions (e.g., college access, employment and career choices, and family formation). NELS does include cognitive testing (not necessarily assessment), but these measures have been used as predictors more frequently than they have been used as outcomes.

To some extent, the constituents of NAEP and NELS have become less distinct during the 1980's. As NAEP began to collect more student background and school characteristic data (frequently using NELS items) and NELS (HS&B) was used for analyses of school effects, program effects, and other studies employing the cognitive test data (frequently using NAEP items), many users of one system began to use both. Many users have insatiable appetites for data. And it is true that NAEP and NELS are improved when they provide more comprehensive data from school principals, teachers, parents, students, and school records. In addition, many analysts interested in school effects would like to have the extensive achievement testing from NAEP included in NELS:88. Hence, one component of the interest in merging NAEP and NELS is the hope that a merged super study will provide more comprehensive data than either current study.

Part of the interest in merging NAEP and NELS stems from the data burden issue. In 1988, both NAEP and NELS will be collecting data from students in schools offering eighth grade. In 1992, both NAEP and NELS will collect data from students in schools offering grade 12. Many State and local school staff feel that two studies is at least one too many.

Finally, some of the interest in merging NAEP and NELS is the desire to reduce cost. It is hoped that the cost of a merged NAEP and NELS would be less than the total costs of separate NAEP and NELS.

The major impediment to the merger of NAEP and NELS is fear of loss. The fear is well-founded. Without careful planning a merged NAEP/NELS study could destroy 20 years of trend data, discourage or prevent the growth of theories or models of many behaviors and transitions, limit the utility or comprehensiveness of achievement report cards, and destroy the user-friendliness of the data. Finally, NAEP and NELS include special, distinct components. For example, NAEP has conducted assessments of adult literacy and NELS has been used to study postsecondary access, choice, and persistence.

Any proposal for merging NAEP and NELS should be considered in light of these issues. As a set of indices, or constraints, a super study should be judged against the following:

1. Preservation of Trend Comparisons. The data should allow comparisons with NAEP data, and the earlier NELS studies.

2. Comprehensiveness. The data should include student achievement data as comprehensive as NAEP provides, and student background data as comprehensive as NELS provides. In addition, data from school principals, teachers, parents, and school records should be included.

3. Burden Reduction. The respondent burden should be less than the sum of burdens of NAEP and NELS(:88).

4. **Cost Reduction.** The cost should be less than the costs for separate NAEP and NELS.

Implications

The four constraints on merger logically imply several important design limits on a super study. To preserve trend comparisons, the cross-sectional components of the super study should gather data at about grades 4, 8, and 12. The longitudinal component should gather data in grades 8, 10, and 12 (and beyond). Clearly, the cross-sectional and longitudinal components overlap in grades 8 and 12. The burden reduction and cost reduction constraints imply that only one data collection should be conducted in these overlapping grades. The comprehensiveness constraint implies that the grade 8 and 12 data collection should have testing as broad as NAEP and student characteristics data as broad as NELS.

The cross-sectional component of the super study should preserve the grade 4 data, and it is possible to begin the longitudinal study in grade 4 as well. However, longitudinal components are very expensive. Since most fourth grade students stay in school through the eighth grade, cross-sectional studies can provide nearly the same growth curve information that a longitudinal study would provide. The cost reduction constraint implies that too little is gained from a fourth grade longitudinal component to add it to a fourth grade cross-sectional component.

Timing. The preservation of trend comparisons and cost constraints suggest that timing is a very important consideration for a super study. NAEP plans extend through 1990 and NELS plans extend through 1994. To modify either would cause problems that might violate either the trend or cost constraints. Hence, a logical start date for a super study is sometime after 1990, probably in 1992 or 1994.

Merged Super Study Design

The following design describes one possible super study. It satisfies all four constraints. For discussion, the study has been described as beginning in 1992, but it could work just as well beginning in 1994.

1992

Cross-sectional component. The data collection should employ samples of 4th, 8th, and 12th graders. The instrumentation and procedures for the 4th and 12th graders should be as similar to NAEP as possible. The instrumentation for 8th graders should be enhanced in the following ways:

1. The number of BIBed or matrix sampled assessment items should be as comprehensive as NAEP, but structured to yield useful scores for each student. This may require fewer BIB packages and more students, but the two necessary attributes are common scores for each student in the sample and continued NAEP comprehensiveness.

2. Tracing/directory information should be collected from each student. This data must flow to the data collector for processing. The directory data should be separated from other data, encoded for confidentiality purposes, and securely stored for future use.

3. Student characteristics and a few items known to be precursors for later transitions should be uniformly collected from each student. Current NAEP procedures use matrix sampling techniques for these items, so the individiual burden would increase.

4. School principals, teachers, and parents should complete questionnaires.

Longitudinal component. With enhancements 2 and 3, the 8th grade sample in the cross-sectional component will serve as the base-year for longitudinal follow-up. Hence, separate samples for cross-sectional and longitudinal studies are not needed, with corresponding savings in costs and respondent burden.

1994

Cross-sectional component. The 1994 cross-sectional component of the super study should be a typical NAEP in grades 4, 8, and 12, without any of the enhancements included in the 1992 study.

Longitudinal component. The 1994 follow-up of students from the 1992 8th grades should include the following:

1. Selection of 10th grade schools attended by the 1992 8th graders in a fashion that allows group administration of instruments. School effects can be a priority, with policy relevant sub-groups to students oversampled. The tracing/directory data from 1992.2 is fundamental to identifying the 10th grade schools. In addition, identification of policy relevant subgroups will require data from 1992.3 and possibly 1992.4.

2. The cognitive test measures used in 1994 should be a subset of the 1992 measures--geared to providing estimates of growth on the common scores generated in 1992. The tests will be shorter and require less administration time.

3. All dropouts during the 1992 to 1994 period should be followed.

4. The sample should be freshened to represent all 10th graders in 1994. (This will allow comparisons with HS&B and NELS:88.) The freshening sample should be as small as possible.

5. School principal and teacher data should be collected.

6. Offerings and enrollment data should be gathered.

1996

Cross-sectional component. Fourth and eighth grade data should be collected in 1996 in a fashion as similar to NAEP as possible. The 12th grade data will be collected in the longitudinal component.

Longitudinal component. The 1992 8th graders, freshened with 10th graders in 1994, should be followed in 1996. The limits of this component are as follows:

1. Select 12th grade schools attended by the 1992 8th graders in a fashion that allows group administration of instruments. School effects can be a priority, with policy relevant sub-groups to students oversampled. This suggests that many of the same schools identified in 1994.1 will be included in 1996. The tracing/directory data from 1992.2 is fundamental to identifying the 12th grade schools. In addition, identification of policy relevant subgroups will require data from 1992.3 and possibly 1992.4.

2. For the tests, the number of BIBed or matrix sampled assessment items should be as comprehensive as NAEP,

but structured to yield useful scores for each student. This may require fewer BIB packages, but the two necessary attributes are common scores for each student in the sample and continued NAEP comprehensiveness. Additional test items should be included as necessary to measure cognitive growth.

3. The sample should be freshened to represent all 12th graders in 1996. (This will allow comparisons with NAEP, NLS-72, HS&B, and NELS:88.) The freshening may require substantially more students than were used in 1994, to accommodate the comprehensive NAEP tests.

4. All dropouts (and stopouts) during 1992 to 1996 should be in the sample.

5. School principals, teachers, and parents should complete questionnaires.

6. High school transcripts should be collected and coded.

1998

Cross-sectional component. The 1998 cross-sectional component should be a standard 4-8-12 NAEP.

Longitudinal component. A subsample of the 1992 8th graders, including dropouts, should be followed using mail survey techniques. This survey will focus on postsecondary access and choice, student financial aid, and employment issues.

Summary

Merger of NAEP and NELS is possible in a fashion that will maintain capacity for trend comparisons, comprehensive national report card data on achievement, and theoretical developments related to student transitions while reducing respondent burden and total costs. Possibilities, however, require careful planning before they can be converted to studies. With recent advances in testing, computer technology, and data collection, it may be possible to merge NAEP and NELS (at the 8th grade) as soon as 1992.

The distinct activities of NAEP (e.g., adult literacy) and NELS (postsecondary access and choice) should be continued using separate studies. The merged design may be repeated as funding constraints and managerial considerations allow. The merged super study should reduce the total costs (probably by about the cost of one longitudinal follow-up), but the management complexity will increase.

Finally, the merger issues may be vitiated if NELS returns to its original focus on postsecondary topics. It is doubtful that this is a real possibility, but it deserves consideration. In any case, a merged super study will require careful planning and substantially more management than either study currently receives. This paper has not discussed the mechanics of grants and contracts, but these may be cause for substantial effort before a super study can begin. Experience with NELS suggests that if a super study begins in 1992, the Center for Education Statistics staff should begin working on the specifications as soon as possible.

Part B

Comments from the NAEP Perspective

Gary W. Phillips, Chief,
Cross Sectional and Special Studies Branch

Introduction

The desire to combine NAEP and NELS is motivated primarily by the hope that such a marriage would improve the understanding of relationships between teachers, schools, and student outcomes. Such improvements would provide the Nation with a better data base for trend monitoring, policy analysis, and applied research.

The idea of integrating NAEP and NELS has strong intuitive appeal. Such a merger suggests reduced data collection burden on schools and students as well as cost saving. Assuming that continuity with the past can be maintained (through bridge studies), a merger would potentially capitalize on the strengths of both assessments. For example, the extensive background data base on the student's school and home environment is a major strength of NELS. The major strength of NAEP is its capacity to provide more comprehensive data on student achievement than any other national survey. NELS is more limited in this area.

It should be noted, however, that a full integration of NAEP and NELS is not necessary to achieve the goals mentioned above. A strong effort to coordinate the two surveys should be more than sufficient to allow data from one to be compared with the other. By coordinating the surveys, they would be permitted to achieve their independent objectives, maintain continuity with the past, and yet provide data that can be linked at some unit of analysis. The following proposal suggests a design that would provide data that could be paired or matched at the school level in a national sample.

Coordinating NAEP AND NELS

NAEP has been designed to monitor long-term trends in student achievement across a broad range of subject matter. It has always focused on outcome skills and knowledge that a consensus of professionals agree are important to the Nation. Since NAEP is an assessment survey, it has been designed to answer questions in the aggregate about how much U.S. students know and do. Since its purpose has been descriptive, and not explanatory, matrix-sampling of items and an overall cross-sectional design have yielded adequate estimates of national proficiency. The desire to make NAEP answer more policy-relevant questions has required the assessment to include background items related to the home and school environment. However, it is widely recognized that the estimation of the effects of these types of variables will be attenuated in cross-sectional data. A

longitudinal design is required to measure the cumulative, interaction and unique effects of input and contextual variables. NELS is better designed to address these issues.

For the most part, NELS does not focus on curriculum and instructional outcomes, nor is its primary purpose to report on what the Nation's students know and do. The primary purpose of the survey is to describe and explain student transition patterns through the American education system. Faithful to its purpose, NELS employs a longitudinal design which allows the effects of exogenous variables to be estimated using the student as the unit of analysis.

There are at least two ways of coordinating NAEP and NELS without radically altering the structure of either. The first would be through using common data elements in each respective data collection instrument. The second approach would be through a coordinated sampling of schools. Either approach would represent a step forward in integrating the two surveys. Combining the two approaches would substantially improve the usefulness of both NAEP and NELS.

Coordinated Instrument Design

The obvious approach to improving the comparisons between NAEP and NELS is to share data elements. This can be done more often, even under the current BIB version of the multiple matrix sampling design used by NAEP. This will be especially true when NAEP begins using a variation on the BIB design in 1988. At that time NAEP will no longer give items from multiple subject areas to the same student. Instead each student will be administered enough items within the same subject area to estimate a proficiency score. The change, along with a common block of background questions, will result in a rectangular data matrix for each subject area. This improvement in NAEP will provide a more useful data base for secondary users and facilitate comparison between NAEP and NELS.

Advances in testing technology may make it possible to go beyond the mere sharing of common items between NAEP and NELS. Recent developments in item-response theory models and Bock's "duplex design" could permit both NAEP and NELS to use the same instrument design (at least for achievement items). Such an instrument would yield aggregate scores on the proficiency of students within a large number of objectives (for NAEP), and at the same time provide reliable estimates of individual performance across objectives (for NELS). The following represents a straightforward procedure that could be used to accomplish this task.

1. Within each of 4 subject areas an assessment instrument would be developed with items distributed across forms according to the duplex design.

2. In NAEP a 30-item form could cover 30 objectives across 3 broad domains within each subject area. This would provide about 10 items per domain per form. This design would provide NAEP 4 scores per student per subject area (3 domains and 1 total test score), and an additional 30 aggregated scores per subject area (1 for each objective). Administering tests in 4 subject areas, and allowing 30 minutes per test, would require 120 minutes of achievement testing.

3. In NELS a 20-item form could cover 20 objectives across 2 broad domains within each subject area (these objectives would be subsets of the NAEP set). This design would provide NELS with 3 scores per student per subject area (2 domains and 1 total test score). Administering 4 achievement tests and allowing 20 minutes per test would require 80 minutes of achievement testing.

4. The design should work if at least 30 students were tested within each school, and each student received a randomly replicated version of each of the 4 subject area forms.

5. A final requirement of the design is that items appearing in the test would be chosen randomly from a bank of items representing each of the objectives. In addition, all items in the bank would need to have been previously calibrated and equated using item response theory methods.

Sharing more common data elements, or using the same instrument would be useful ways of relating NAEP to NELS. However, this approach has several shortcomings which can only be remedied by additional design enhancements. For example, NAEP items have previously been used in NELS studies as well as in the Second International Math Study. In each case, differences in sampling, time of testing, test administration, and other context effects have questioned the credibility of comparisons. In addition, sharing common data elements only permits the comparison of descriptive statistics and does not provide for the estimation of relational indices (such as correlation coefficients) between surveys. The calculation of relational indices would require that NAEP and NELS data be paired at some unit of analysis.

Coordinated School Sampling

The goal of the sampling design should be to permit the results of NAEP and NELS to be paired at the school level in a nationally representative sample. There are at least two ways this goal might be accomplished. The first is to coordinate the sampling of NAEP and NELS such that each is collected from a

subsample of schools. The sample of students within each school would be randomly divided into a NAEP sample and a NELS sample. Although this design does not increase the data collection burden on any individual student, it does substantially increase the burden across students within the subsample of schools.

An alternative design would involve coordinating the sampling of NAEP and NELS such that the samples would contain no common schools. However, the school sampling would be conducted in such a way that NAEP and NELS would contain a subsample of matched schools. The matching would be done based on important school characteristics.

Matching schools has many desirable features. It represents a useful coordination of the independent surveys without attempting a complete merger. It would build a bridge between the two studies by providing merged data sets down to the school as the unit of analysis. It would permit NAEP to continue providing a national report card, while NELS focuses on growth and change. Since the unit of analysis is the school, the data obtained from the subsample should be sufficient to assess school and program effects. Furthermore, the subsample would yield unprecedented data on the correlations between achievement and school characteristics. Finally, the matching procedure holds the promise of providing these data without any increase in burden or costs.

Final Comments

This design is an effort to coordinate NAEP and NELS in such a way that data points from each survey could be paired at the school level. Such a linkage would permit applied researchers and statistical reporting agencies to correlate data obtained by NAEP with that obtained by NELS in a nationally representative sample. Furthermore, using the school as the unit of analysis should be sufficient to provide policy makers with useful information.

Another important benefit of the above design is that it allows NAEP to be responsive to other national agendas. For example, the same method of coordinated sampling could link NAEP (and NELS) to the Elementary/Secondary Integrated Data System (ESIDS). Coordinating these three major national surveys in this way would fulfill the promise of the comprehensive and integrated survey system, called the Redesign Project within the Center for Education Statistics (CES). In addition, coordinated sampling is consistent with the recent recommendations of the study group commissioned by Secretary William Bennett to suggest ways of improving NAEP. The report of the study group suggested that NAEP use student cohort sampling, and be linked to other national surveys. The design discussed in this paper obtains data based on student cohort sampling from NELS, and links NAEP, NELS and ESIDS at the school level in a nationally representative sample.

PART C

Comments on the Conference

David A. Sweet, Director
Education Outcomes Division

Much of the interest in merging the national assessment and the longitudinal studies program is focused on 1988. In that year NAEP will be assessing 8th grade students (as well as 4th and 12th graders and the traditional 9-, 13-, and 17-year-olds) and NELS:88 will launch its base year data collection at the 8th grade level.

Several speakers at the conference have already explained, however, that the national assessment and longitudinal studies data collection operations cannot be merged into a single super study in the near future, e.g., 1988. Although the instrumentation, sampling and field operations can and definitely will be coordinated in 1988, a full scale merger in 1988 is precluded by time schedules for consensus development and planning as well as existing contractual and grant agreements, and, in the case of NAEP, policies on data confidentiality and congressional specifications on governance.

In the intermediate or long run, however, all of these strictures could conceivably change. The issue then becomes whether a merger of some sort at some time in the future might not only be possible but also be desirable. Phrased this way and temporarily holding off the question, "How soon?" the issue of whether to merge or not merge becomes more intriguing.

Calvin Jones has done a marvelous job of discussing NAEP and NELS:88 similarities and differences and suggesting ways of merging the two in the future. Joan Baron and Pat Forgione have not only described a state perspective of State information needs, possible National/State data system linkages, and data burden considerations, but also set forth some models for an integrated system of cross sectional and longitudinal data collection at the National level. Bruce Spencer discussed sampling problems in merging cross sectional and longitudinal studies. And Alan Ginsburg et al. described Federal data needs which need to be addressed, many by national assessment and longitudinal studies.

Dennis Carroll was correct in chronicling the NAEP and LSP histories. LSP was initially focused on postsecondary access, choice, and persistence and NAEP was focused on assessment of elementary and secondary aged youth--9-, 13-, and 17-year-olds. LSP was recognized as the best national data on high school students' transitions into postsecondary education and the world of work and, by 1980, also the best information on high school

students, their schools, teachers and parents. NAEP was recognized as the best nationally representative data on trends in cognitive achievement.

The two programs have been made more similar over time. LSP has moved to lower grades--from a 12th grade base year in 1972, to a 10th grade base year in 1980, and now an 8th grade base year in 1988. NAEP has added grade samples, added student background and classroom process items, and moved to summary scale score reporting of cognitive achievement.

With all these changes in the two programs, this is an appropriate opportunity to take stock of where we have been, where we are now, and where we might best make adjustments if prudent. Clearly, two uncoordinated programs are undesirable. There are a myriad of options, however, between total integration of the programs and planned differentiation and linkages of two programs. In the coming year we would like to develop criteria for examing the alternatives. Dennis Carroll has suggested four: trends, coverage, burden, and cost. I would like to elaborate on trends preservation.

Increasingly those most familiar with data of the two programs point to the value of trends. For NAEP, these are simple trends at the three specified age levels. For LSP, these are the somewhat more sophisticated trends in transition rates (e.g., changes over time in the access to postsecondary education either overall or for special subgroups of interest to policy makers and researchers). Thus, those who have used the data most have asked that whatever changes might be made in either program, please do not destroy trend information, or the potential for trend analyses.

I personally find this argument for preservation of important trend data thoroughly convincing. The practical consequence of a commitment to trends information is that priority be given to a high standard of comparability in instrumentation content, sample coverage, and other factors which might destroy comparability such as changes in procedures, time schedules, or context.

The particulars of where longitudinal components might best be built onto the national assessment programs or where the assessment program might best be expanded to overlap with longitudinal study activities and interest are far from clear. Conference presenters have offered a variety of approaches. Budget considerations alone, however, will force us to be selective.

Strong arguments have been put forward for longitudinal studies starting at all sorts of age or grade levels--i.e., starting as late as first time students in postsecondary education and as early as birth.

No criteria were were offered for choosing among these alternatives, in part I believe because the criteria that have been used in the past conflict with one another and have different meanings at different times. These are the criteria:

- o addresses policy issues upon which there is little credible data,
- o provides trend information or some other basis for drawing inferences on the most pressing issues,
- o addresses the information needs of several audiences, and
- o fits into an overall system of data collection.

Of course, there have not been many new longitudinal studies and LSP has never gone below the 8th grade. So no trends would be at issue for a new longitudinal study starting at a lower grade level - e.g., a longitudinal study springing off of the 4th grade NAEP sample.

Current interest in longitudinal studies can be seen as falling into three areas, each with its own sets of policy issues. NELS:88 plans address two of the three areas--secondary and postsecondary but not preschool/elementary. The NELS:88 8th grade cohort will examine the precursor conditions that lead to dropping out of school and the policies, practices, and conditions that lead to school effectiveness and student growth. The second area is addressed by the NELS:88 postsecondary cohort, which if funded will be linked with the National Postsecondary Student Aid Study. This cohort will be used to examine postsecondary persistence and choice and the impact of student financial aid. The third area is less developed and therefore is not included in NELS:88. It has been talked about starting with 5-year-olds, kindergarteners, or 1st graders and follow-ups extending at least to the 4th grade.

The national assessment program activities coincide with these longitudinal study plans and interests only at the 8th grade. How assessment might profitably be extended to these other levels will require further discussion and will be covered in two other conferences in this series.

I would like to set aside the other alternatives until this conference series is completed, and focus for the moment on a merger starting at the 8th grade level. Several of the conference presenters, and Dennis Carroll and Gary Phillips in their comments, have discussed how an 8th grade national assessment and an 8th grade longitudinal study might be linked at some future time. The years 1992 or 1994 have been suggested.

It is not unrealistic to assume that by 1992 or 1994 the Nation's experience with Darrell Bock's duplex design and other assessment designs will have advanced to a stage that we can have the breadth of subject matter coverage needed for assessments and the uniformity of student scores needed for longitudinal studies examining growth in achievement.

We even may have a two-stage adaptive test like the ones now being examined in the Illinois and California assessment programs. Without major developmental work, however, I assume full computer adaptive testing will not yet be available for studies starting in 1992 or 1994.

Length of data collection sessions remains an unresolved issue. Some discussants initially argued for the NAEP 1 hour assessment while others preferred the 3 hours of questionnaires and testing used in the LSP. No one offered a way that 3 hours of information could be collected in 1 hour. There is some hope that computer adaptive testing and questionnaire administration can shorten the 3-hour sessions below the 2-hour mark. All agreed that the special value of the LSP is in the rich questionnaire data and the special value of national assessment lies in the fuller assessment of subject areas.

The second follow-up of NELS:88 will occur in 1992. At that time most of the 8th graders will be high school seniors. To avoid excessive burden on the schools, therefore, it seems advisable to hold off initiating a new longitudinal study until 1994. Budget and staff resource limitation considerations would also point to 1994 rather than 1992.

Working backwards that means a field test would be conducted in 1993, and a contract for field operations and data processing would be let in 1992 and written in 1991. That means that a planning conference would be held in 1990. Commissioned papers for the planning conference should therefore be initiated early in 1990.

In the interim we will sift through the guidance we receive from this and the other two conferences in this series. We will also be looking at ways that the national assessment and longitudinal studies program activities might be coordinated and linked (in both the short and long run) with other Center data collection programs like the School and Staffing Survey of schools and teachers.

More Bang for the Buck: An Integrated Data Collection Strategy

Alan Ginsberg and Valena Plisko, U.S. Department of Education
Nora Guhl and David Myers, Decision Resources, Inc.

Federal data collection activities in education are undergoing extensive review. One of the key questions facing policymakers is the nature of the blueprint for collection of data involving comparisons of students' educational performance over time and the related variables that affect this performance.

Repeated measures of student educational performance over time, either longitudinal (collecting data from the same student) or cross-sectional (collecting data from different students), serve several important functions for educational policymaking:

- Over-time data are important for monitoring trends in student educational performance. For example, since 1970 the National Assessment of Educational Progress (NAEP), a cross-sectional survey, has provided repeated measures of student performance in specific subject areas, at three grade levels.

- Over-time data are also important for assessing the short-and long-term effects of educational programs and family characteristics on student performance. For example, High School and Beyond (HSB), a longitudinal survey of high school sophomores and seniors, has been used to examine the relative effectiveness of public and Catholic schools in fostering growth in achievement from sophomore to senior year (Hoffer, Greeley, and Coleman 1985). With retrospective data on mother's employment status during the student's childhood, the HSB survey has also been used to examine the long-term effects of mother's work on the student's achievement (Milne, Myers, Rosenthal, and Ginsburg 1986).

Repeated cross-sectional and longitudinal data can yield important policy-relevant information about educational outcomes and processes. Both types of data, however, are expensive to collect from nationally representative samples. Experience with these data is now extensive, and strategies for improving their effectiveness warrant consideration. The trade-offs and optimal uses of repeated cross-sectional surveys and longitudinal surveys require explicit and careful examination. Thus far, repeated cross-sectional and longitudinal data collections have been assessed individually rather than as part of an integrated data collection strategy.

This paper lays the groundwork for such an integrated assessment along four lines of inquiry. First, we consider opportunities for better use of extant data bases. Second, we

review what we have learned about improving the measurement of variables and we identify a new area of concern. Third, we consider attempts based on both cross-sectional and longitudinal data to make causal inferences about program effects. Finally, we raise issues that educational policymakers should address in the effort to develop an integrated data collection strategy.

Using Extant Data Bases

Collecting nationally representative data is costly, and funds for collecting statistics are limited. The Center for Education Statistics' budget in fiscal year 1987 is $14.1 million, 17 percent below its budget of 1980 (adjusted for inflation). Obviously, we must expand the use of data that have already been assembled. But the Federal Government does not always organize its statistical collections in ways that promote the efficient use of the information it produces. When educational researchers work with large nationally representative data bases, they tend to use only those collections that deal exclusively with education. But major national surveys conducted outside the education field can often yield additional insights into important policy questions of educational concern. Moreover, these surveys could be enhanced by the selective addition of questions related to educational experiences and outcomes.

Department of Education data bases. The major Department of Education data bases that provide over-time information on student educational performance are listed in Table 1. About $65 million has been spent thus far on these surveys. Researchers in the education field have primarily used the longitudinal data bases (i.e., HS&B and NLS-72) produced by the Center for Education Statistics. These data bases have been supported by ready access to user tapes accompanied by clear and well-documented user manuals. This availability and ease of use has produced a rich variety of studies on a broad range of school, student, and family topics. The SES, SIMS, and NAEP data bases, in contrast, have had little application outside the federally funded contractor studies, so their potential has been restricted. For example, the Educational Testing Service has not fulfilled its commitment to develop a comprehensive user tape of NAEP data, which was a major feature of the contract award. As a result, few third-party analyses of NAEP data have been conducted.

Data bases outside the Department of Education. Panel B of Table 1 lists several data bases that have been produced outside the Department of Education. These data bases provide over-time data on education-related topics. Although their primary focus may not have been education, these data bases, both cross-sectional and longitudinal, provide information on educational processes and outcomes. More than $100 million has been invested in producing these data bases thus far.

These major data bases prepared by Federal agencies other than the Department of Education remain an undertapped resource for investigating educational concerns. These data bases can be profitably used to analyze educational issues, to check results from analyses of Department of Education data bases, and to aid in the design of future surveys. For example, educational researchers could use the CPS household survey to analyze trends in enrollment or educational attainment in preschool or private educational institutions. Researchers could use the MF data to explore the changing attitudes and behaviors of youths, including sex education and attitudes, drug use and school drug policies, and student education goals and expectations. And the researchers could use SIPP, with its detailed household income information, to examine how student and parent financing for postsecondary education changes over time as Federal and other student financial aid opportunities change.

Case study. The untapped potential of existing data bases funded outside the Department of Education to inform educational issues can be illustrated with the Labor Department's NLSY. Although this rich source of information on youth experiences has been tapped by economists, education analysts have been slow to recognize its potential. It can be used, however, to explore a major educational concern--the different estimates of the national high school dropout rate. One common method of calculating this rate uses school district or state administrative records to compute the graduation rate as the ratio of students who enter the 9th grade to the number of graduates four years later. Nationally, this calculation yields a graduation rate of 71 percent and a residual dropout rate of about 29 percent (Office of Planning, Budget, and Evaluation 1987). One obvious problem with this measure of the dropout rate is that it is not derived from the educational paths of individual students; hence it is only an implicit measure of the dropout rate.

Longitudinal surveys that track individual students have the potential to improve knowledge of dropout processes and estimates of dropout rates. The best currently available educational data base, HSB, however, begins surveying only with students at the end of their sophomore year, thereby missing an unknown number of early dropouts. In this respect, the Labor Department's NLSY is advantageous because it covers people who were 14 to 16 years of age in 1979. This age cohort is generally within the legal age for mandatory schooling, and hence it has few members who have dropped out of school. Their schooling status from 1979 to 1984 is described in Table 2.

The proportion who, in any particular year, have left school without obtaining a diploma or GED is shown to rise from 4.4 percent in the first year to 15.9 percent in 1983, and then to decline slightly. These rates are consistent with the Census Bureau's estimated dropout rate for 16- to 23-year-olds, i.e., about 14 percent (Bureau of the Census, 1985). However, the

cumulative percentage of persons who have, at some time, dropped out before completing high school reaches 22 percent, about one-third more than the "current" dropout rate in any one year. This cumulative rate of ever dropping out is consistent with estimates obtained using administrative records (i.e., the graduation rate). Table 2 also shows that many dropouts (about one-third) eventually return to school, accounting for the lower final dropout rate measured in the Census Bureau's questionnaires.

This example clearly shows the overlap between the concerns of educational researchers and policymakers and information collected by Federal agencies other than the Department of Education.

Improving the Measurement of Educational Processes and Outcomes

Obtaining nationally representative samples that allow for the intensive statistical analysis of subgroups has been a primary objective of most large-scale data collection efforts. Hence, large surveys have developed elaborate procedures to minimize sampling error. For example, NAEP uses a variant of matrix sampling called Balanced Incomplete Block Spiralling. See NAEP (1985) for a description of this procedure. An equally important source of error--measurement error--also requires explicit attention. In order to reap the full benefits from nationally representative samples, we need to have reliable and valid measures of educational processes and outcomes. Reliable measures are simply those that yield consistent results across repeated measurements (Carmines and Zeller, 1979). Valid measures are those that accurately measure the phenomenon in which we are interested (Carmines et al., 1979). To achieve reliable and valid measures that minimize measurement error, we need to pay careful attention to the appropriateness of the measures we choose and the respondents from whom we obtain information. Three examples illustrate the importance of obtaining reliable and valid measures of educational processes and outcomes: reliability of responses, validity of outcome measures, and validity of measures of educational processes.

Improving the reliability of responses. The major Department of Education data bases differ in the choice of respondents who provide information. For example, to obtain information about parents' education, some surveys ask the parents, others ask the student, and some ask both parents and students. The choice of the student, the parent, or the school to provide information has implications for the accuracy of the information.

The choice of the respondent often reflects the primary purpose of the survey. For example, the NAEP instrument was developed mainly as a survey to assess student outcomes and trends. NAEP represented a landmark effort in the early 1970's when it provided a common national assessment of student

performance on basic academic subjects. Any information it obtained on student backgrounds or school processes was somewhat incidental to the basic NAEP mission. Thus, NAEP has relied heavily on student self-reports to derive information on student educational and family backgrounds.

By contrast, longitudinal data bases have traditionally emphasized understanding the schooling and home processes that generate the observed outcomes. Longitudinal survey efforts have shown that student reports are often poor measures of family income. For example, the correlation between student and parent reports of parents' income is .3 in the HSB survey (Rosenthal, Myers, Milne, Ellman, and Ginsburg, 1983). Also, student self-reports of grades are frequently found to be inflated when compared with school transcripts (National Center for Education Statistics 1984). In response to these sources of error, the Department of Education's longitudinal survey collected school transcript data for student grade and course information and obtained family background statistics directly from parents. The NAEP survey has not yet built such features into its regular data collection plan.

Improving the validity of outcome measures. The validity of measures of educational outcomes is another area of concern. For example, if we want to measure reading comprehension, we need to develop a measure that accurately reflects that skill. The empirical measure should be a good proxy for the cognitive skill of interest.

If empirical measures obtained from various surveys are tapping the same cognitive skill, then the same or similar cognitive skills should exhibit similar long-run trends. We can illustrate the application of a criterion of intersurvey consistency in trends over time with respect to reading scores of high school seniors over time. Despite all the care given to developing the NAEP outcome measures, there are unexplained differences between NAEP trends in high school reading performance and trends exhibited by other major standardized tests. Accurate estimation of these trends has become increasingly important as reports have concluded that high school student performance has declined over the last 15 years and as States and school districts have responded by introducing educational reform, especially in high schools.

As Table 3 shows, the NAEP reading score trends fail to support the prevailing view and, in fact, between 1971 and 1984 NAEP reading scores for 17-year-olds increased by one-tenth of a standard deviation, or approximately 4 percentile points. Most of the increase has occurred since 1980, but the scores exhibited a slight increase during the 1970's. If correct, the NAEP evidence fails to corroborate a serious decline in test score among high school students in reading performance.

Improving the validity of measures of educational processes. To understand educational outcomes, we need to have valid measures of the educational processes that generate those outcomes. For example, we have long sought to identify the school factors that affect student achievement. Large-scale national studies have tended to use easily measured variables, such as teacher experience, teacher education, and other measures of school resources as proxies for the qualitative differences in educational processes that we know exist among classrooms. But we have obtained few consistent results about the effect of school factors on achievement, as measured by these gross indicators of the learning process in classrooms. Hence we apparently need to refine our measures of the classroom learning process (i.e., make them more valid) in order to understand why some students learn more than others.

One way to improve measures of educational process is to collect data on program content and processes. The International Association for the Evaluation of Educational Achievement's math survey, SIMS, has demonstrated the feasibility of collecting educational process data from large samples. SIMS collected detailed information on the characteristics of schools, teachers, and mathematics programs; on classroom process, such as teachers' allocation of time, assessment of class ability, classroom organization, use of resources, instructional techniques and goals, and beliefs about effective teaching; and on students' background, including parents' education and occupation and students' time spent on homework and attitudes towards math. The SIMS study also collected information on student exposure in the classroom to concepts covered in the math tests. The first round of analyses of these rich data has uncovered a variety of instructional and curricular practices (e.g., excessive of topics and dominance of arithmetic in the junior high curriculum) that are associated with the lower math achievement of U.S. students compared with their European and Japanese counterparts (McKnight, Dossey, Kifer, Swafford, Travers, and Cooney, 1987).

Making Inferences About Causality

Causal inferences are made in program evaluation research and in more basic research that attempts to uncover fundamental processes that influence educational outcomes. Ideally, researchers need experimental data (i.e., data generated from a randomized design) to make causal inferences. However, in the social sciences, we almost always deal with nonexperimental data. For example, often when estimating the impact of participation in an education program, researchers find that students have been assigned to a treatment group on the basis of need (e.g., low-achievement test scores) rather than on random assignment. Thus, the effect of the program on achievement cannot be distinguished from the effect of the low-achievement that caused the children to be placed in the program.

Estimation of the relationship between academic tracking and students' academic performance provides another example in which the use of nonexperimental data makes it difficult to obtain "pure" (unbiased) estimates of the effect of the program. Students are not randomly assigned to academic tracks. Students in an academic track are generally more motivated toward school and tend to have higher achievement-test scores than students in other tracks. Thus, the effects on achievement of motivation and of tracking cannot be distinguished.

In nearly all analyses that examine questions such as the impact of educational programs on academic performance there is an implicit temporal ordering of events. Specifically, program intervention is followed by measurement of academic performance. Cross-sectional data can provide information to examine the correlation among variables, but this information is often insufficient for making causal inferences. To understand the causal sequence it is necessary to draw on data that capture the dynamics of the ordering process. Data such as these can be obtained from the inclusion of items on survey instruments that refer to prior events (retrospective data) or through the multiple administration of survey instruments to the same persons over a long period of time (longitudinal data).

Using longitudinal data. Given the prevalance of nonexperimental data, researchers generally attempt to control for observable and unobservable characteristics that differentiate the groups under study and relate to the outcome. In recent years it has become apparent that the use of longitudinal data with the use of statistical controls may provide greater leverage in teasing out the effects of program participation on selected outcomes.

The potential advantages of longitudinal data over cross-sectional data can be illustrated by past attempts to identify the effects of the Federal Government's Chapter 1 compensatory education program on educationally disadvantaged participants. Researchers have tried to approximate randomized designs by statistically controlling for differences between treatment and control groups with varying degrees of success. Under one approach, researchers used NAEP's repeated cross-sectional data to draw conclusions about the impact of Title I on students' reading achievement (NAEP, 1981). The mean reading achievement of students in schools eligible for Title I services was compared with the achievement of students in other schools at three time points: 1971, 1975, and 1980. The implied treatment or program that was hypothesized to affect student reading achievement was attendance at a school eligible for Title I services. The study concluded that "the overall pattern of a narrowing gap (in achievement) for most population groups at all ages suggests that students in Title I schools are improving at a faster rate than students in non-Title I schools." (NAEP, 1981.) The recent congressionally mandated study of Chapter 1 argues that "it is difficult to estimate the extent to which the patterns of

improvement indicated by NAEP derive from the particular children who received Chapter 1 services." (Kennedy, Birman, and Demaline, 1986:11.) The Kennedy et al. study concluded that the NAEP data were inadequate to support any conclusion about the specific effect of Chapter 1 services on student achievement because the NAEP data did not include pre- and posttest information on actual Chapter 1 participants.

Better evidence about the effect of Chapter 1 services on achievement comes from an analysis of the SES data, an individual-level longitudinal data set (Carter, 1984). Analyses of these data used measures of individual students' achievement before and after participation in Chapter 1 programs, as well as a number of variables measuring student and family characteristics to control for nonrandom selection into the program that would affect postprogram scores. This analysis concludes that there are modest short-term gains in the achievement of disadvantaged students, but not long-term effects. The effects of Chapter 1 were also shown to be greatest among these participants who were least disadvantaged.

In the SES analyses longitudinal data permitted an assessment of pre- and postprogram achievement as well as a means by which biases from nonrandom selection into the program could be statistically adjusted. Although longitudinal data may permit analysts to control statistically for nonrandom selection in the program in some cases (Appendix A contains an example), longitudinal data do not guarantee that reasonable results will be obtained in all cases. Only under certain conditions can longitudinal data provide the information needed to obtain estimates of the program effects. For example, Bassi (1984) found that it was possible to obtain unbiased estimates of the effect of participation in Comprehensive Employment and Training Act (CETA) programs on earnings for some subgroups (e.g., white women) and not for others (e.g., white men). Estimates of CETA's effect on white male earnings ranged from a positive $151 to minus $2,403. The effect of nonrandom selection was so strong that longitudinal estimators were not adequate for purging the effect for some subgroups. This example suggests that large scale longitudinal data may in some cases have to be supplemented with smaller samples of more carefully selected treatment and control groups. Knowledge of the selection process used to sort students into treatment and control groups could provide the leverage necessary for estimating the impact of educational programs.

Using retrospective data. Longitudinal data have shown that some programs may have long-term consequences that may not be adequately measured immediately upon the conclusion of the program. For example, longitudinal data on the Perry Preschool project in Ypsilanti, Michigan have shown that early schooling has long term effects on a variety of behaviors, including truancy, use of special education services, high school grade point average, dropping out of school, employment, transfer

payments, pregnancy, and delinquency (Berrueta-Clement, Schweinhart, Barnet, Epstein, & Weikert 1984, Gramlich 1986).

Because of the time and cost involved in collecting longitudinal data as well as the analytical problems engendered by sample attrition, it is useful to explore the conditions under which long-term effects may also be captured with retrospective data. Retrospective data can be obtained from administrative records, such as student transcripts, or from respondents. Research has shown that transcript information is a much more reliable indicator of student course-taking patterns than student self-reports (Campbell, Orth, & Seitz 1981).

When recall is likely to be accurate, however, retrospective data obtained from respondents may provide a cost-effective method for obtaining over-time data. For example, in interviews of parents, HSB collected retrospective information on mother's employment status during several stages of the student's childhood. On the basis of this retrospective information, analyses of HSB data have shown that early maternal employment may have long-term negative effects on youths' achievement (Milne et al., 1986).

Given the usefulness of retrospective data as well as the relatively low cost, it is important to learn more about the accuracy of retrospective data obtained from respondents.

Conclusion

With funding for data collection by the Department of Education likely to be limited, ensuring efficiency assumes great importance. Policymakers should draw on their two decades of experience with such collections to consider strategies for improving the policy and research payoffs.

The current political climate also is especially conducive to reform. There is extraordinary interest in educational data now as a result of the many reports chronicling the problems with the U.S. educational system. Policymakers are asking why we only learned about the problems of our schools so long after they occurred. In this climate the Department of Education is reassessing its data collection efforts. Reorganization of its activities has brought NAEP under the jurisdiction of the Center for Education Statistics; this move enhances the possibility of coordinating NAEP with other data collection efforts.

This paper suggests a set of strategies for improving one important type of survey--national collections of data on individual students. Recommendations concern three main areas: improving data utilization; enhancing the quality of information; and enhancing the capacity of national surveys to identify causal relationships through expanding longitudinal data collections.

Improving data utilization. The first step is to encourage greater application of data bases produced outside the Department of Education to educational problems. The Department of Education can support the use of "outside" data bases by funding studies that extend analysis beyond Department of Education data or by encouraging participation in education conferences by experts who are knowledgeable about "outside" data bases. Department of Education staff also needs to work with staff from other Federal departments such as the Department of Labor when they are designing questionnaires. In this way Department of Education staff can make known its needs and determine whether the instruments used by other departments may be an efficient means for obtaining certain educationally relevant data.

The second step is to prepare user tapes with clear and detailed documentation for users. Where these have been prepared, such as for the major longitudinal surveys from the Education and Labor Departments, literally hundreds of studies have been carried out. This is in stark contrast to the NAEP and SES which, despite their $50 million+ cost, have gone virtually unused by researchers other than the federally funded contractors who collected the data. Small, competitively awarded Federal grants also could increase the yield derived from multimillion-dollar investments.

Enhancing the quality of information. It is important for the Department of Education to exercise greater quality control over the reliability and validity of measures. National studies are especially prone to suffer problems from unreliable data because of large samples. Where feasible, the Department of Education should periodically check data it collects for consistency with other sources of similar data. Also, the Department of Education should devote greater attention to collection of information from the appropriate respondent. In the case of student's family background, this may necessitate the added expense of a household survey.

It is also important to obtain more information on the schooling or home processes that give rise to the observed outcomes. Specialized surveys such as NAEP, which historically focused primarily on obtaining outcome information, could be made more useful if supplemented with questions about the courses students take as well as their content. The International Association for the Evaluation of Educational Achievement's math survey, SIMS, which explored the "opportunity to learn" as a predictor of student achievement, serves as an important model of the ways in which information about course content can be obtained.

Enhancing the capacity of national surveys to identify causal relationships through expanding longitudinal data collections. Cross-sectional studies such as NAEP should be supplemented with smaller subsamples of students who are followed

longitudinally. The subsample could represent all students or follow a particular population of interest, such as students from a particular grade or racial group. Although somewhat more complicated, this integrated cross-sectional and longitudinal approach appears statistically feasible (Bock, 1986; Jones, 1986; Spencer, 1986).

Retrospective information should be collected on respondents when reliability is satisfactory. Retrospective data cost relatively little time and money to collect compared with true longitudinal data. Administrative records, such as transcripts, afford a relatively accurate means for obtaining a student's academic history. Respondent recall may yield adequate responses for some kinds of information of high policy interest, such as mother's work or marriage history. At a minimum, the Department of Education should fund studies of the accuracy of different types of retrospective information.

In closing, although our discussion has centered on the improvement of large-scale national data collections, we have recommended that small-scale studies may be appropriate supplements or alternatives to large-scale studies in a number of cases. Our overriding concern is that we have research tools that enable us to address the issues that face educational policymakers. Given the paramount importance of learning about the effects of educational programs as well as the concern about cost, educational policymakers may have to consider greater use of small-scale, detailed case studies.

TABLE 1

Selected List of Extant Data Bases

Panel A—Department of Education Data Bases

Data Base	Year Started	Sample	Sample Size	Cost[a]	Focus
High School and Beyond (HSB)	1980	Nationally representative sample of high school sophomores and seniors	58,000	$15 million	Educational achievement, high school experiences, transitions to postsecondary school, employment, marriage, and parenthood
National Longitudinal Study of the High School Class of 1972 (NLS72)	1972	Nationally representative sample of high school seniors	24,000	$11 million	Transitions to postsecondary school, employment, marriage, and parenthood
National Assessment of Educational Progress (NAEP)	1969	National representative sample of 9-, 13-, and 17-year olds, and young adults ages 26 to 35	75,000 to 100,000 young persons have been assessed annually or biennially	$20 million+	Achievement in reading, writing, mathematics, and science, social studies, and other learning areas taught in schools
Sustaining Effects Study of Title I (SES)	1976	Nationally representative sample of students in grades 1 to 6	120,000	$17 million	Effect of Title I services student achievement
Second International Mathematics Study (U.S.) (SIMS)[b]	1981	Nationally representative sample of 8th- and 12th-graders	12,000	$3 million	Content of math curriculum, methods by which math is taught, math achievement

(continued)

Panel B--Data Bases Supported Outside the Department of Education

Data Base	Year Started	Sample	Sample Size	Cost[a]	Focus
National Longitudinal Survey of Labor Market Experience of Youth (NLSY)	1979	Nationally representative sample of youth aged 14 to 21 years.	12,686	$30 million to date	Formation of human capital, labor market experience, transition to adulthood
Monitoring the Future (MF)	1975	Nationally representative sample of high school seniors.	3,600 annually	$2 million annually	Student drug use and activities
Panel Study of Income Dynamics (PSID)	1968	Nationally representative sample of families	3,000 initially 5,000 currently	$22 million to date	Determinants of family income and its change
Survey of Income and Program Participation (SIPP)	1983	Nationally representative sample of households	20,000 initially; 12,000 currently	$60 million plus	Participation in federal programs
Current Population Survey--Education Supplement (CPS)	1968	Nationally representative sample of households.	60,000	October 1986 education supplement cost an additional $160,000.	Educational enrollment and attainment

a All cost figures are approximate. Cost figures are not adjusted for inflation.

b SIMS was jointly funded by the U.S. Department of Education and the National Science Foundation.

TABLE 2

Dropout Rates Derived from NLSY Data, 14- to 16-Year-Old Cohort (in 1979)

	1979	1980	1981	1982	1983	1984
Age (Years)	14-16	15-17	16-18	17-19	18-20	19-21
Current dropout [a]	4.4	7.8	12.7	15.3	15.9	15.4
New dropout [b]	4.4	4.3	6.0	4.4	2.1	1.0
Cumulative ever dropout [c]	4.4	8.7	14.7	19.1	21.2	22.2
Percentage of new dropouts who had reentered school or who had obtained a diploma or GED by 1984	30.2	38.1	27.6	32.1	27.2	NA

[a] Current dropout is the percentage of youths in the age group who are not enrolled in school and have no diploma or GED.

[b] New dropout is the percentage of youths who drop out of high school for the first time in each survey year.

[c] Cumulative ever dropout is the cumulative percentage of youths who are "ever dropouts." This percentage approximates the true sample percentage of the 14- to 16-year-old cohort who ever drop out. The percentage of new dropouts in each survey year cannot be strictly added, because the denominators differ in each year due to sample attrition. Checks, however, show that the approximation is quite close. Another estimate of the cumulative percentage of ever dropouts was obtained by dividing the weighted sum of new dropouts in all years by the weighted sample size in 1979. This estimate, 22.1 percent, represents a lower bound of the cumulative percentage of "ever dropouts" since sample attrition between 1979 and 1984 (about 4 percent in this cohort) may have reduced the total number of respondents who were "ever dropouts."

TABLE 3

Change in Reading Test Scores of High School Seniors, 1971-84

Year	SAT	ACT	ITED	NAEP[a]	NLS72-HSB
1971-75	-0.19	-0.15[b]	-0.09	+0.01	--
1972-80	--	--	--	--	-0.21
1975-80	-0.09	+0.04	-0.14	0	--
1980-84	+0.02	+0.04	+0.11	+0.09	--
Total change	-0.26	-0.07	-0.12	+0.10	-0.21

Sources: The College Board, 1986; ACT, 1986; R. Forsyth, 1986; NAEP, 1985; Rock et al., 1985.

Notes: Change is reported in standard deviation units. SAT verbal, ACT English, and Iowa Test of Educational Development (ITED) "interpretation of literary materials" test scores are reported. Test scores are assigned the "spring date" of the school year. For example, a test that is administered in the fall of the 1975-76 school year is labeled 1976. This labeling convention produces some discrepancy with published sources.

[a] NAEP test-takers are 17 years old and may be in any grade.

[b] 1970-75.

Appendix A

Suppose we have a situation in which it is hypothesized that prior achievement, a set of exogenous observed variables such as family income and parental educational attainment, and an unobserved individual specific factor affect achievement at time t. Further, imagine that a compensatory education program is administered to a select group of students between time t and t+1. It is hypothesized that the same variables affect achievement at each measurement period. Such a process may be captured be the following statistical equations:

(1) $Y_{it} = a_t Y_{it-1} + X_i B_t + I_i + e_{it}$, and

(2) $Y_{it+1} = a_{t+1} Y_{it} + X_i B_{t+1} + cP_i + I_i + e_{it+1}$

where Y refers to student i's achievement (i=1,...N) at time t(t=1,...,T); X is a vector of observed charactertistics such as family income and parental educational attainment; I is the unobserved individual specific factor for student i that is constant over time; e is a random error; P indicates whether a student received services between time t and t+1; and a, B, and c are conformable vectors to be estimated. The coefficient c is an estimate of the impact of participation in a compensatory education program on Y (e.g., student's achievement). To obtain an unbiased estimate of c requires knowledge of prior achievement, the observed characteristics, and the unobserved characteristics. Although the last component cannot be measured, it is possible to secure an estimate of c by taking a first difference. That is, when equation (1) is subtracted from equation (2), the following is obtained:

(3) $Y_{it+1} - Y_{it} = a_{t+1} Y_{it} - a_t Y_{t-1} + cP_i + (e_{it+1} - e_{it})$, or

(4) $Y_{it+1} = A_{t+1} Y_{it} - a_t Y_{t-1} + cP_i + e^*_{it+1}$

where $A_{t+1} = (a_{t+1} + 1)$ and $e^*_{it+1} = (e_{it+1} - e_{it})$.

Equation (4) shows that by taking the difference of equations (2) and (1), we are able to cancel out the effects of the unobserved individual specific factors. Equation (4) can be estimated, for example, with ordinary least squares.

References

ACT. National trend data for students who take the ACT assessment. Iowa City: ACT, 1986.

Bassi, L. Estimating the effect of training programs with non-random selection. Review of Economics and Statistics 1984, 56: 36-43.

Berrueta-Clement, J., Schweinhart, J., Barnett, W.S., Epstein, A., and Weikart, D. Changed Lives: The Effects of the Perry Preschool Program on Youths through Age 19. Ypsilanti, Michigan: The High/Scope Press, 1984.

Bock, D. Instrument design for a combined NAEP and NELS. Paper presented at the Center for Education Statistics conference on "The National Assessment of Educational Progress (NAEP) and the Longitudinal Studies Program (LSP)--Together or Apart?" December 11, 1986.

Bureau of the Census. Current Population Reports, Series P-20, No. 400, 1985.

Campbell, P., Orth, M., and Seitz, P. Patterns of Participation in Secondary Vocational Education. Columbus: National Center for Research in Vocational Education, The Ohio State University, 1981.

Carmines, E., and Zeller, R. Reliability and Validity Assessment. Beverly Hills: Sage Publications, 1979.

Carter, L. The Sustaining Effects Study and elementary education. Educational Researcher, August/September 1984, 4-13.

The College Board. College-bound seniors. New York: The College Board, 1986.

Congressional Budget Office. Trends in Educational Achievement. Washington: CBO, 1986.

Forsyth, R. (Iowa Testing Program), personal communications. October 1986. Mean ITED test scores by grade and subtest for the state of Iowa. Iowa City: Iowa Testing Programs, undated and unpublished tabulations.

Gramlich, E. Evaluation of education projects: The case of the Perry Preschool Program. Economics of Education Review, 1986, 5, (1):17-24.

Heckman, J., and Robb, R. Alternative methods for evaluating the impact of interventions: An overview. University of Chicago mimeo, 1985.

Hoffer, T., Greeley, A., and Coleman, J. Achievement growth in public and Catholic schools. Sociology of Education, 1985, 58(April):74-97.

Jones, C. How to optimize and articulate a longitudinal and a cross sectional research program. Paper presented at the Center for Education Statistics conference on "The National Assessment of Educational Progress (NAEP) and the Longitudinal Studies Program (LSP)--Together or Apart?" December 11, 1986.

Kennedy, M., Birman, B., and Demaline, R. The Effectiveness of Chapter 1 Services. Washington: U.S. Department of Education, 1986.

McKnight, C., Crosswhite, J.F., Dossey, J., Kifer, E., Swafford, J., Travers, K., and Cooney, T. The Underachieving Curriculum: Assessing U.S. School Mathematics from an International Perspective. Champaign, IL: Stipes Publishing Co, 1987.

Milne, A., Myers, D., Rosenthal, A., and Ginsburg, A. Single parents, working mothers, and the educational achievement of school children. Sociology of Education, 1986, 59:125-39.

National Assessment of Educational Progress. The Reading Report Card, Progress Toward Excellence in our Schools: Trends in Reading over Four National Assessments, 1971-1984. Princeton: Educational Testing Service, 1985.

_____. Has Title I improved education for disadvantaged students? Evidence from three national assessments of reading. Denver: NAEP, 1981.

National Center for Education Statistics. Quality of Responses of High School Students to Questionnaire Items. NCES 84-216. Washington, D.C.: NCES, 1984.

Office of Planning, Budget, and Evaluation (U.S. Department of Education). State Education Statistics Supplement: Student Performance and Resource Inputs, 1985 and 1986. Washington: OPBE, 1987.

Rock, D., Ekstron, R., Goertz, M., Hilton, T., and Pollack, J. Factors Associated with the Decline of Test Scores of High School Seniors, 1972 to 1980. Washington: Center for Statistics, U.S. Department of Education, 1985.

Rosenthal, A., Myers, D., Milne, A., Ellman, F., and Ginsburg, A. The failure of student-parent cross-validation in the High School and Beyond Survey. Proceedings of the Social Sciences Section of the American Statistical Association Annual Meetings, 1983.

Spencer, B. Sampling problems in merging a cross-sectional and a longitudinal program. Paper presented at the Center for Education Statistics conference on "The National Assessment of Educational Progress (NAEP) and the Longitudinal Studies Program (LSP)--Together or Apart?" December 11, 1986.

Shooting at a Moving Target: Merging the National Assessment of Educational Progress and the Longitudinal Studies Program--A State Perspective

Joan Boykoff Baron and Pascal D. Forgione Jr.
Connecticut State Department of Education
and
Marc Moss
Harvard University

Introduction

At the outset, we'd like to begin with two reactions. The first is that we feel that it is vitally important, at this time, as we prepare to implement the next wave of two large national testing programs to explore possible ways to merge them. The emergence and importance of developing state indicators for reporting on the impact of recently enacted and implemented state reform and improvement efforts is responsive to increased demands for public accountability. Second, the position reflected in this paper is only one State's reaction. Connecticut's views are a function of our priorities and our own data collections procedures. We have not had time to canvas our colleagues in other States and we do not claim to represent a consensual viewpoint.

The Moving Target: The Need for a Reconceptualization

If we ever felt like we were shooting at a moving target, it was in writing this paper. Between December 4, when we wrote our first draft of this paper and December 9 when the Study Group on the National Assessment of Student Achievement met, several critical events took place. In response to these events, we changed our conception on the merge from one composed of two or three separate assessments operating independently, yet in concert, to one large assessment with an expanded set of purposes (see Elliott and Hall, 1985 and Selden, 1986 for background).

On December 4, we mailed to Washington a draft of this paper which represented an incremental approach to merging aspects of the new NAEP with the upcoming NELS:88. The paper supported the recommendation of the Council for Chief State School Officers (CCSSO) which called for a national achievement indicator and comparable state statistics related to three areas: 1) Resources; 2) Processes and 3) Outputs (Selden, 1986) that could be reported state-by-state. We wrote that paper from the perspective of a State which values valid and reliable information about input and output indicators of educational equity and excellence. Toward those ends, we had already made a commitment to participate in the CCSSO testing program (at that time projected for the winter of either 1988 or 1989 and possibly using the NAEP item bank). We had also decided to piggyback on

the NELS:88 by administering also in the winter of 1988, the requisite achievement and questionnaire batteries to 8th grade students, their parents, their teachers, and their principals in 50 Connecticut schools.

Our decisions to participate in these national assessments were made incrementally. However, in reflecting on these commitments, we became concerned about the administrative burdens that this configuration of assessments might place upon our schools. This is illustrated in Table 1.

TABLE 1

HYPOTHETICAL CONNECTICUT STATE TESTING SCHEDULE
WITH INDEPENDENT NAEP AND NELS

TEST	DATE
CONNECTICUT MASTERY TESTS (GRADES 4, 6, AND 8).../	OCTOBER 1988
CCSSO TEST (USING NAEP GRADES)...../	JANUARY 1988
NELS:88 (GRADE 8)...../	FEBRUARY 1988
NAEP (GRADES 3, 7, and 11 or 4, 8, and 12)...../	APRIL 1988
CAEP (GRADES, CONTENT DOMAINS, AND TESTING DATES TO BE DETERMINED)	

The Connecticut Context

In order to more fully appreciate the administrative burdens required by these additional tests, it is important to understand the Connecticut testing context. In 1983-84, partly in response to a second wave of concern for educational equity, Commissioner Tirozzi secured legislative approval for the development of the Connecticut Mastery Testing program. This fall, the Department fully implemented a criterion referenced testing program in grades 4, 6, and 8 in reading, language arts, including writing, and mathematics. This program entails seven to eight hours of testing for each child in those grades during the months of September and October. Each student takes 2 hours of testing in reading comprehension; 1 in listening comprehension/spelling/notetaking; 2-3 in mathematics; 1 in writing mechanics/study skills; and 1 in a direct measure of writing. Results are returned to school districts in December, aggregated by school and classroom. Two copies of the test results are included for each student--one for the student's parents(s) or guardian(s), and the other for the student's record.

Our second testing program, the Connecticut Assessment of Educational Progress (CAEP), beginning in 1971 was administered in grades 4 (winter), 8 (fall), and 11 (spring) in order to make our CAEP results comparable with NAEP results. However, in 1984, with the advent of Mastery Testing in grades 4, 6, and 8, the focus of the CAEP program was shifted temporarily to the secondary school. In 1983-84, we conducted our first assessment in a vocational educational area (i.e., business and office education in the areas of accounting, general office, and secretary). In 1986-87, we will conduct our second assessment in a vocational education area--this time in the industrial arts and technology areas of drafting, graphic arts and small engines. This year, we will also conduct our first foreign language assessment for all students in grades 9-12 taking at least the second course in French, German, Italian, Latin and Spanish. We are currently postponing a redesign of our CAEP program until final decisions are made related to the timing of testing in NAEP, NELS:88, and the CCSSO program. (For a more complete description of our testing program, see Tirozzi, Baron, Forgione and Rindone [1985].)

The Recommendations of the Study Group on the National Assessment of Student Achievement

On Thursday, December 5, the Study Group on the National Assessment of Student Achievement, chaired by Lamar Alexander, met in Chicago to make recommendations regarding the future of NAEP. These recommendations were based upon a comprehensive set of commissioned papers and a series of meetings conducted by the national study group. It was recommended that the Nation's Report Card (the new name for NAEP) be a two-year cycle of four content areas assessed in the winter at grade 4, 8, and 12 in all

50 States, and include an out-of-school 17-year-old cohort study. In addition, the national study group supported the notion of merging NELS and NAEP.

This new framework provides an efficient and suitable framework for coordinating and potentially merging the three separate national testing programs (i.e., NAEP, NELS:88 and the CCSSO). This would be accomplished by selecting a sufficiently large number of students from each state to accurately generalize too the 50 separate student populations.

Both of the designs depicted in Tables 2A and 2B would permit a longitudinal design. The advantage of the 2B design is that the time interval between the beginning of the administration of testing to new cohorts is decreased and that now timely state/national reports can be provided. For example, grade 8 and grade 12 reports can be produced every four years and transition from school to work (year 14) will be also available every four years. Thus, some information can be reported every other year as part of a state or national report card. In addition, the linking of NELS with NAEP will allow NELS to focus on the in-depth background et al. variables that have been too superficially measured previously under NAEP.

One of the advantages of the proposed merger would be to address a possible confound that might exist in a longitudinal study which assesses the same children every two years. Known as "the Hawthorne effect", this repeated testing could indeed influence the behavior of these students. With the present design we can began to isolate the effects of this confound by comparing the students in the longitudinal study with the others in the same grades.

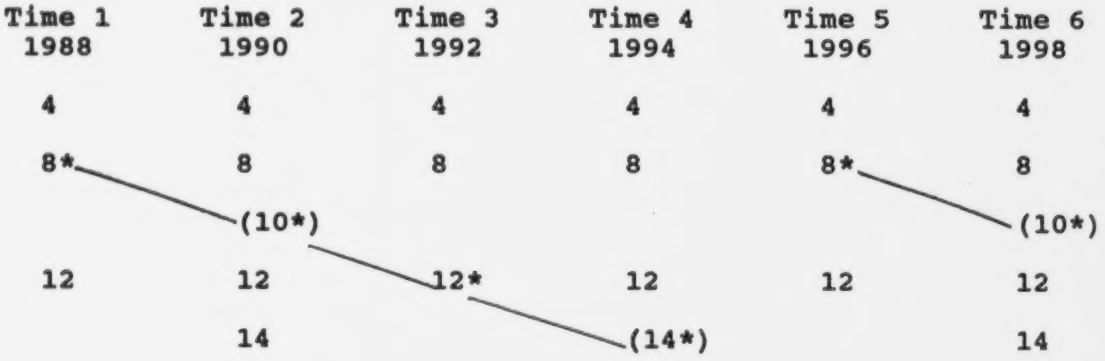
The newly conceived national assessment is primarily designed to provide longitudinal data on the progress of the Nation and its States rather than on the achievement of individual students, classes, schools or districts. However, it will be possible for school districts and schools to "piggyback" on NAEP (a position we would strongly endorse) for those jurisdictions that wish to underwrite an enhanced sample.

The Stakeholders

In considering the advantages of a merger, it is helpful to view it from the perspectives of the various stakeholders-i.e, the larger three groups which have a direct involvement with the new assessment (See Table 3). Viewed this way, it appears that there are many more advantages than disadvantages, a point to which we will return later.

TABLE 2A

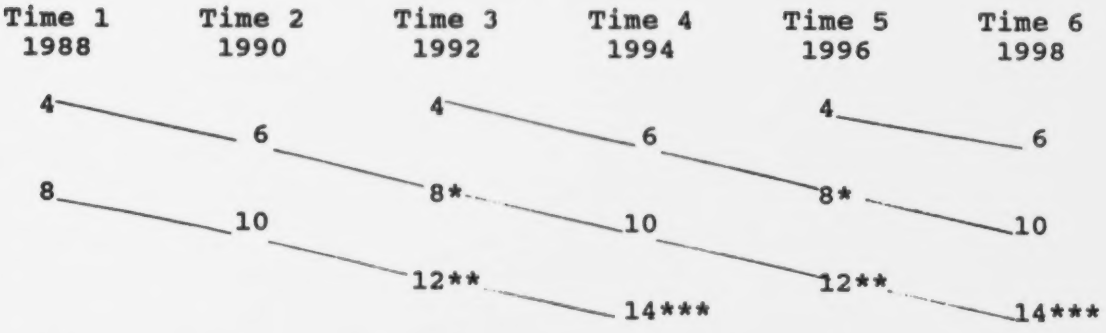
One Possible Testing Schedules for a Merger of NAEP, NELS, and CCSSO*



* and ( ) depict the coordination of NELS data collection with the NAEP assessment cycle.

TABLE 2B

A Second Possible Testing Schedules for a Merger of NAEP, NELS, and CCSSO*



Many of these ideas are derived from the models described by Jones (1986).

* Means a grade 8 report covering a cohort for four or more years.
** Means a grade 12 report covering a cohort for four or more years.
***Means a grade 14 (transition from school to work) report covering a cohort for six or more years.

Table 3

The Advantages and Disadvantages of the Merger for Various Stakeholders

STAKEHOLDER	ADVANTAGES	DISADVANTAGES
Federal	More "bang for the bucks" (i.e., cost efficiency)	Possible difficulty in making direct comparisons over time due to changes in sampling plans
State	State-to-Nation comparisons at no cost to the States State-to-State comparisons at no cost to the States	
Districts Schools Classrooms Students	Less burdensome in that fewer interruptions would occur during school year (It is not yet clear whether these tests will have any clear instructional implications. The former NAEP studies had great breadth of coverage but did not report by objective. The recent scales developed by the NAEP in reading and writing described a continuum of performance along a single dimension, but did not tie results to curriculum and instruction)	

The key policy question is what effect will the increased efficiency of a merger have on the overall effectiveness of the programs when contrasted with keeping them separate?

In effect, this will be determined by the level of funding of the merged study. Will the funding sources view a merger as a cost-saving mechanism and thereby trim the budget since one large assessment will probably cost less than two smaller assessments? Or, will they see the merger as an opportunity to assess the status of education in our Nation more comprehensively using the full funding that has been appropriated for both programs?

An Enhanced National Report Card

Connecticut supports the latter approach and advocates an enhancement of the merger. We have three suggestions as to how to best use these funds. First, the focus could be expanded beyond what was possible in the two separate studies. If sufficient funding were to be available, an enhanced National Report Card could include a new component missing from both NELS:88 and NAEP--that of classroom observations. This would enable us to move beyond performance data and self-report data to classroom process data. Both the NAEP and NELS:88 studies, as presently conceptualized, would benefit from visits into classrooms. Heretofore, both studies have attempted to learn what happens inside classrooms by asking students, teachers, and principals series of multiple-choice questions on questionnaires about the frequencies with which various activities take place. Our own studies in CAEP have provided us with evidence that one learns precious little from questionnaires about what really happens in classrooms. We are not suggesting that the respondents are dishonest. Rather, the questionnaire items and responses are too insensitive to the instructional process to detect meaningful differences. Furthermore, people are not always conscious of their metacognitive strategies and specific aspects of their environment so they construct plausible but inaccurate scripts (Nisbett & Wilson, 1977). Visits to classrooms in specially selected schools have the potential to provide insights into why students in some schools drop out of school at a higher rate than students in other schools.

A second use of funds to improve upon present studies might be a systematic attempt to understand the causal patterns which underlie the relationships observed in the achievement and questionnaire data. Consider what the benefits are of bringing together, in planning the studies, the Nation's leading researchers, theorists, and psychometricians to design a study that would enable us to explore causality. This might require a combination of causal modeling, cross-lagged panel designs, time series analyses, etc. It might even generate some new statistical approaches. But, if done a priori, we could collect the appropriate data that would give us insight into the factors or processes both mediating and moderating expected causal relationships rather than trying to determine causal relationships after the fact, from correlational data (see Baron, 1980 and Baron & Kenny, 1986).

A third possible addition to a National Report Card would be to probe the depths of item response theory to achieve some of the ends envisioned in the Bock and Mislevy (1986) paper on duplex design employing vertical and horizontal calibrations.

Undoubtedly, it seems optimal to design a matrix sampling approach that permits analysis of both individual and program data. However, as the authors point out, it may be difficult to satisfy the assumptions of linearly ordered content domains (LOCDs) and the need for the factor loadings in the general and specific factors of items in a given element to be in a constant ratio. An enhanced National Report Card Study would allow us to explore alternative designs for data collection and analysis to bring our reality closer to our vision.

Summary

In reviewing the original purposes of the two programs it appears that none of them is sacrificed by a merger. Furthermore, a merger enables the addition of two new purposes to obtain comparisons (see Table 4).

Table 4

Purposes of Testing Achieved by NAEP and NES:88 Kept Separate Versus Merged

Original Purposes of Testing Programs	Separate NAEP	Separate NELS	MERGER
to monitor student achievement over time (i.e., cross-sectional data; e.g., Grade 4 vs. Grade 4)	X	X	X
to determine curricular strengths and weaknesses	X		X
to promote curricular advancement	X		X
to raise standards of achievement	X		X
to understand the relationship between achievement and other variables	X	X	X
to follow a longitudinal cohort over time (e.g., Grade 8 to Grade 10 to Grade 12)		X	X
New Purposes of Testing			
to compare state-to-nation			X
to compare state-to-state			X

In closing, Connecticut's support of a merger hinges on the assumption that a reduction of testing and disruption would be obtained by interfacing NAEP, NELS:88 and CCSSO. The realization of this end depends on its technical feasibility. It assumes that we can combine the two samples and that we will be able to link future data to past data. It is critically important that a merger not interfere with our Nation's ability to chart its educational progress over time. Once a technical plan is developed to accomplish a merger without sacrificing longitudinal data, it is important that representatives from the States review and comment upon it.

REFERENCES

Baron J. B. (1980) How school achievement and attitude toward school are influenced by a set of demographic, ecological, and psychological variables: a causal model analysis. Ph.D. dissertation. The University of Connecticut.

Baron, R. M. & Kenny, D.A. (1986) The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology (in press, 51, 6).

Bock, R. D. and Mislevy, R., (1986) Comprehensive educational assessment for the states: the duplex design. (Mimeographed paper)

Elliott, E. J. and Hall, R. (1985) Indicators of performance: measuring the educators. Educational Measurement: Issues and Practice 4 (2) 6-8.

Jones C. C., (1986) Relationships between the National Assessment of Educational Progress and the National Educational Longitudinal Studies Program. (Mimeographed paper)

Nisbett, R.E. and Wilson, T.D. (1977) Telling more than we know: Verbal reports on mental processes. Psychological Review 84, 231-259.

Selden, R. (1986) Recommendations for State-by-State Assessment of Student Achievement. Presented to: Committee on Coordinating Educational Information and Research Council of Chief State School Officers. October 17, 1986.

Tirozzi, G., Baron, J. B., Forgione, P. D., Rindone, D. A. (1985) How testing is changing education in Connecticut. Educational Measurement: Issues and Practice 4 (2) 12-16.

How To Optimize and Articulate a Longitudinal and a Cross Sectional Research Program

Calvin C. Jones
Senior Survey Director
NORC

Introduction

The optimization of a research design is the process of combining multiple, possibly inconsistent goals with realistic assessments of operational constraints into a compromise design that one hopes will succeed, at least partially, in satisfying the original research objectives. Optimization is frequently considered to be a mechanism for coping with constraints that preclude the use of preferred means of pursuing objectives, and is often considered equivalent to what Herbert Simon referred to as "satisficing."

For example, faced with a fixed budget and the need to conduct a variety of policy studies, a research administrator has relatively few choices for allocating resources. Objectives may be prioritized, and only those rated above a threshold value pursued, abandoning objectives with low priority scores. Or, if sacrificing any objective is unacceptable and there are recurring opportunities to choose among them, projects may be scheduled so that all objectives are pursued, but less frequently than might be desirable to meet research or policy goals. Finally, the administrator may attempt to optimize the allocation of resources across projects by combining complementary objectives of two or more usually separate projects into a single effort. Once the specific objectives are identified, and their relative values and costs are established, formal analyses may be carried out to identify the approach that promises the highest return across all goals, given the known constraints.

More generally, however, the techniques of optimization may be used to evaluate whether the objectives of one or more research projects or programs might be more effectively or more economically achieved by reorganizing their design and/or operational features. A by-product of this sort of analysis may be the discovery that reorganization offers opportunities for achieving important new objectives that were previously beyond the reach of the individual projects, because of cost or other operational constraints. However, such discoveries do not typically emerge as miraculous surprises from a formal analysis. First, as in all research design efforts, the opportunities must be identified through an unscientific creative process. Then, the formal techniques can be employed to confirm or disprove objectively what one hopes or suspects to be true.

The title of this paper suggests a more general treatment of the topic of research program integration than I have actually attempted. Rather than discuss theoretical relationships between cross-sectional and longitudinal studies, I have focused specifically on the opportunities and obstacles to integrating the National Assessment of Educational Progress (NAEP) and future cohorts of the National Education Longitudinal Studies (NELS) program. I have tried to approach the problem from a general, systems-analysis perspective, and to make specific recommendations wherever possible. However, at this time, formal mathematical modelling would be premature.

Although the primary features of the NELS program are relatively well-established, for the past several months, the Department of Education has been engaged in a thorough review of NAEP's objectives and design, the result of which may be the introduction of a number of major changes in its basic purpose and operation. Many of the recommended changes, if adopted, would radically alter the desirability and feasibility of specific integrating options. Serious optimization analysis should therefore be delayed until the Department has determined the type and structure of the data it wants from NAEP.

One important exception to that conclusion is that the Department may decide that it wants to introduce the collection of individual-level longitudinal data to its approach to assessment research by re-assessing the same individuals as they progress through elementary and secondary school. While the merits of this idea seem to me to be worth considering, to my knowledge this decision has not been made. Changing NAEP to include longitudinal measurement of student performance would require dramatic shifts away from the traditional assessment design. I have included a brief discussion of this and other more radical possibilities for program integration in a closing section of the paper, but I have not considered changes of this magnitude to be under consideration at this time.

In the absence of clearly specified goals for an integrated longitudinal--assessment program, I have concentrated on identifying the design features of both studies that seem to offer the greatest potential research gains from greater articulation. Space does not permit more than a cursory review of the possibilities, so I have focused the discussion of opportunities for integration around the similarities and differences between the two programs as they now exist, resisting the temptation to suggest entirely new designs. Even the most radical alternative of a fully merged design for collecting longitudinal and assessment data retains most of the basic characteristics of the separate studies. To structure the discussion further, I have used the assumptions and simplifications described below.

1. Assumptions and Simplifications

1.1 The Need To Expand the Use of NAEP and NELS for Education Policy Formation and Analysis

First, I have considered program integration only from the point of view of raising the utility of NAEP and NELS data for applied research and policy formation, and have generally ignored the potential uses of each program for basic research on human learning. In general, this motive focuses more on the limitations of NAEP than those of NELS studies.

1.2 The Need for Historical Continuity in NAEP and NELS

Second, I consider the status quo to be one of the most powerful inhibitors of greater integration of the designs. Both the NAEP and the NELS programs have been operating (quite independently of each other) for 15 to 20 years. Because each program includes a time-series orientation, historical continuity is a crucial feature of the data production methodology and the approach to data utilization. Because major change to either program may seriously disrupt that continuity, any consideration of different approaches to integrating the programs and different degrees of articulation must highlight the difficult choices that the Center for Education Statistics must make about the relative importance of different types of data it seeks, and the allocation of its limited resources to each type.

The inescapable truth is that the more integration is adopted, the more likely that the continuity of one or both of the programs will be damaged. This situation is entirely different from that of articulating cross-sectional and longitudinal research programs still in their development stage, and promises to be more technically difficult and politically intricate. I have presented two extreme options available for integrating the programs--marginal adjustments in the separate studies versus fully integrating the two studies in a merged design--in part to illustrate the costs and benefits of disrupting the continuity of either program. Throughout the discussion, I have tried to keep in focus the problems of implementing the actual transition from the current operation of either program to any more integrated alternative.

1.3 A Focus on Data Quality and Relative Efficiency of Task Accomplishment Rather than Technical Properties of the Study Designs

Third, since my own background is primarily that of a developer and manager of policy research design and data production, I am interested in the possibilities for operational and procedural articulation between NAEP and NELS, as well for substantive, intellectual and scientific integration. I have

therefore looked for opportunities for program integration that are likely to increase both the quality and the efficiency of the data collection, testing, data processing and analysis tasks within the overall enterprise, and thus have the greatest chance of enhancing the utility of data from both programs for applied research. My aim is to suggest choices for project design and organization that would reduce the duplication of effort in the current arrangement, would provide both studies with more types of data, significantly improved measures, and expanded analysis possibilities, and would offer greater manageability and control over the quality and timeliness of the data production process--while remaining within the same level of resources currently allocated to these studies.

Other papers produced for this conference focus specifically on the more technical aspects of sample design, testing, and data analyses. Although the choices over population definitions and sample design, and the approach to the measurement of developed ability profoundly influence the possibilities for program integration, I will limit my attention to the impact of these decisions upon the desirability and feasibility of integration, leaving the technical issues to sampling statisticians and psychometricians.

1.4 A Focus on Quality Enhancement with Constant Costs Rather than Constant Quality with Reduced Costs

Fourth, I have assumed that the Federal interest in articulating the two programs reflects a desire for improving the quality and power of quantitative data to address increasingly complex questions about the condition and impact of education, and that the primary motive is not primarily to cope with current or anticipated reductions in research funding levels. Thus, I have not considered cost cutting to be a primary goal of integration. While some of the recommendations would indeed lower total costs, many others would have little impact on program costs or would require additional resources to implement. I have concentrated on the problems of optimizing a study design to accommodate diverse and sometimes conflicting research goals, and have considered the only cost constraint to be the need to remain within the levels of effort currently allocated to both studies.

1.5 A Focus on Two Extreme Choices for Integrating the Programs

Of the infinite number of models that might be examined for articulating NELS and NAEP, I have considered only two extreme types:

1. comparatively minor changes to the design of each program, while retaining the separate identities and unique characteristics of each; and

2. completely merging every feasible operation of the two programs into a unified study that attempts to accommodate the full set of objectives pursued by both using a single sample for each cohort studied.

Both types of change may be desirable. In some instances, changes of the first type can provide the basis for the transition to more radical integration. For example, standardizing population definitions and coordinating sampling procedures may be initiated at any time, whether or not longitudinal and assessment data are ever collected from the same individuals. In many respects, however, reaching the second goal may require the development and use of techniques and methods not now employed by either study if a merged design is to succeed. Perhaps most relevant here is the need for a far more efficient approach to cognitive test administration than either study now employs.

Prior to discussing specific proposals for articulating the designs, it may be helpful to describe the main features of the two programs.

2. Brief Description of NAEP and NELS

Both the National Assessment of Educational Progress (NAEP) and the National Education Longitudinal Studies (NELS) programs trace their origins to the late 1960's, a time when the (National) Center for Education Statistics (CES), the primary data gathering and reporting unit within the U.S. Department of Education, was pressed by both the executive and legislative branches to collect data on a national and regional basis that would improve understanding of the dynamic properties of our elementary and secondary school system. At its inception, each of these two programs had its own political and research constituencies marshalling the resources and forging the design and methods to address a particular set of issues.

The unique scientific, methodological, and operational problems facing each of the new enterprises were numerous and formidable. Many talented, dedicated individuals in both the private and public sectors demonstrated enormous creativity in developing solutions appropriate for the times and for the distinctive needs of each study. The result, nearly two decades later, is a pair of highly specialized research programs, each demanding substantial dollar resources on a recurring basis, with relatively little naturally-occurring articulation between them, and with a strong tendency within each to expand and build upon its own roots, and to resist gratuitous or fashionable change.

2.1 The National Education Longitudinal Studies Program

The fundamental purpose of the NELS surveys has been the study of stability and change in the educational activities of American youth, including the extent to which students'

background characteristics and other exogenous variables affect educational achievement, and the subsequent impact of educational activities and attainment on such other outcomes as labor force participation and economic well-being, family formation, military service and citizenship. The program consists of three studies: the National Longitudinal Study of the High School Class of 1972 (NLS-72); High School and Beyond (HS&B); and the National Education Longitudinal Study of 1988 (NELS:88).

NLS-72 began in the fall of the 1971-72 school year with a survey of some 23,000 high school seniors in a nationally representative sample of over 1,000 public and private secondary schools. In school-based sessions, each surveyed student completed a self-administered questionnaire and a cognitive test battery, each requiring about one hour. A School Questionnaire was completed by the principal of each sampled high school, and a Student Record Information Form was completed using the school's cumulative file for each student. After the base year survey, the full sample was resurveyed four more times, and a subsample of 15,000 was selected for the Fifth Followup Survey in 1986.

Although much cross-sectional analysis was carried out using NLS-72 base year data, the main educational focus of the survey was the longitudinal study of access to, choice of, persistence in, and completion of postsecondary educational activity, and the impact of postsecondary education on employment and other activities. In the most recent followup, the National Institute for Child Health and Human Development recognized the unique capability of this cohort to support the study of family dynamics and joined with the Education Department in sponsoring a portion of the study.

HS&B began in the 1979-1980 academic year with even larger samples of two grade cohorts in over 1,000 public and private high schools. The HS&B design expanded upon NLS-72 by including a 10th grade cohort to permit the study of growth during the last two years of high school, and to begin to study the dynamics of dropout behavior; by expanding topic coverage in both the Student and School Questionnaires to include such issues as school discipline and order and student employment; by including oversamples of policy relevant subgroups such as private school students, linguistic and racial minorities, and high-achieving students of low socio-economic status to support the longitudinal analyses of special policy issues. It included samples of students' parents to support the study of financing postsecondary education; and by collecting a wide variety of objective records data for sampled students, such as Scholastic Aptitude Test (SAT) and Armed Services Vocational Aptitude Battery (ASVAB) scores, high school and postsecondary transcripts, and financial aid records. Subsamples of about 12,000 1980 12th-graders and 15,000 1980 10th-graders have been resurveyed three times since the base year (in the spring of 1982, 1984, and 1986).

The third longitudinal study, NELS:88, breaks much new ground. By starting with a national sample of over 26,000 8th grade students in 1,000 public and private schools in the 1987-88 academic year, this study will collect data necessary for the study of the dynamics of the crucial transition into secondary school, including the choice of school and program, the determinants of tracking and ability grouping. As in the prior studies, questionnaire and cognitive tests covering science, mathematics, reading and language skills, and social studies will be completed by sampled students in school-based survey sessions.

By starting with an eighth grade cohort, the population definition for the study includes nearly all students who may eventually fail to complete secondary school, nearly half of whom had already dropped out of school when HS&B selected its sample of 10th graders. In addition to collecting data from students and school principals, the base year study will also survey one parent for each sampled student, greatly improving the power of NELS:88 to disentangle the effects of the school and home environments on student growth during secondary schooling. Finally, the base year survey will collect data from up to two teachers of each sampled student, in the subjects of either science, mathematics, English-language arts or social studies, including teacher ratings of students, basic teacher characteristics, and teaching practices in the classrooms where the NELS:88 students are taught. Sampled students will be resurveyed as 10th graders in 1990 and as 12th graders in 1992, and, according to current plan, at two-year intervals for up to three additional followups. As in HS&B, the study design calls for the collection and merging of many types of student records data, such as school transcripts and test scores.

The NELS program has succeeded in providing highly usable and frequently exploited data for policy studies, and in continuing to evolve its study designs to correct the weaknesses and expand upon the strengths of prior designs. Yet, despite its many successes, the NELS program has not yet fully overcome one of its primary challenges--the precise, accurate measurement and successful analysis of individual cognitive growth across years of schooling. Test batteries used in NLS-72 were devised as single-measure achievement tests, and were most heavily used as control or classification variables, rather than as dependent measures. Extensive analyses of the 10th and 12th grade test scores of the HS&B 1980 sophomore cohort were largely frustrated by what proved to be relative insensitivity of most of the test batteries to the curriculum patterns experienced by the students.

NELS:88 is attempting to improve upon past results partly by replacing HS&B's single test design with an item-overlap test design that should increase somewhat the efficiency of measurement of students' achievement levels at the 8th, 10th, and 12th grades, with a supplementary teacher survey that should provide improved measures of both student characteristics and of the types of material to which students were exposed, and with a supplementary parent survey that should help to separate home

from school effects in the analyses of achievement and growth. However, although these changes are clearly in the right direction, the fact remains that the testing of NELS:88 students will be limited to a total of 75 minutes, less than 20 minutes each for four separate subject areas. While our hopes are high for substantial improvements over past results, this basic limitation on test administration may simply prove too great to overcome with the current approach.

2.2 The National Assessment of Educational Progress

Central to NAEP's research objectives and design is the repeated measurement, at short, regular intervals, of the level of developed ability of elementary and secondary school students in basic subjects such as reading, mathematics, writing, science and social studies. By testing three age (and/or grade) cohorts (most commonly 9-, 13-, and 17-year-olds) in each assessment, data from each NAEP cycle permit analyses of the differences in developed abilities among the three groups. By assessing the same subject areas at intervals, data from the time-series of NAEP surveys permit analyses of trends in what each age cohort has learned or can do, and analyses of changes over time in the differences in performance or capabilities among the age groups measured, and subgroups within each age cohort.

In contrast to NELS studies, because the main purpose of NAEP is to describe population trends rather than to support conventional, individual-level survey analyses, NAEP does not attempt to collect exactly the same data elements from all sample members. By allocating the assessment items to test booklets so that each examinee encounters only a small portion of the total item pool, NAEP collects performance measures on a very large number of test exercises related to a great variety of curriculum objectives, while holding the response burden for each individual student to a very low level (about one standard class period). NAEP is thus capable of reporting student performance measures on an exhaustive collection of single items and composite scales for the student populations sampled.

To accomplish these objectives, the NAEP design must sacrifice certain other goals, most notably the ability to support more than a minimal amount of individual student-level analysis, and virtually no class-, school-, district- or state-level analysis. Since different students are given many different versions of the NAEP forms, only a small minority of the students in each cohort sample provide responses to the same set of assessment exercises or background variables. Indeed, since so little individual analysis can be done, very little background data on students' educational activities and experiences, and the educational support systems in their homes is collected.

Besides restricting data collection procedures, other limits were placed upon the design. For example, selection of the

samples of students within each school for participation in NAEP has been shrouded in such extraordinary secrecy that even the data collection contractor does not retain a record of the sampling materials or the identities of the selected students. Supplementary data collection activities (e.g., obtaining school records or transcripts for NAEP student samples) are thus made difficult at best.

To a great extent, these limitations were designed into the earliest NAEP surveys as a means of reassuring State and local educators as well as other important political constituencies that the purpose of NAEP was not to evaluate specific students, schools or States against national standards, and in particular that NAEP was not a vehicle for establishing a national elementary and secondary school curriculum. The price of State and local participation in a national assessment was a deliberate restriction through design features on the use of NAEP data for broad, multi-purpose policy research. Had this price not been paid, continuing political conflict would likely have prevented NAEP's birth. Until very recently, these design elements were assumed to be so fundamental to the nature of the program that recommendations for major refinements and improvements to the NAEP design have left them generally intact.

Thus, in the context of this paper, NAEP must be understood as a very unusual type of cross-sectional research design in which the primary units of analysis are not the individual responding students (or their parents and teachers), but are several separate but overlapping samples or aggregates of the responding units, each contributing different pieces to estimates of the educational achievement of elementary and secondary school students. This fact complicates the relationship between NAEP and NELS as much as the more apparent differences in research objectives, and must be taken into account under any effort to increase the articulation between them.

3. Opportunities for Integration: Identifying Common Features of NELS and NAEP Surveys

In an earlier position paper, I described several broad dimensions of activity common to NAEP and NELS, identified existing obstacles to articulation, and suggested ways of improving the fit between the design choices made by each. I would like to expand upon the points raised there, and to consider under each point the implications of two possible levels of program integration: design changes that might be implemented in either study separately; and the design implications of a fully merged program. I begin with a general review of the ways that resources are allocated across component activities in order to clarify what types of integration might free sufficient resources to address problems in both assessment and longitudinal research.

3.1 Reducing Duplication of Effort

In recent years, a typical two-year NAEP assessment cycle has required between $8 and $9 million to complete. The cost of NELS cycles varies widely depending upon the presence of supplementary surveys (e.g., parents and teachers) and whether survey administration will occur in schools. Since the first three rounds of NELS:88 will require school based surveys and are likely to include samples of parents and teachers, the average total cost of each of these waves may be similar to those for a NAEP cycle. About half the resources expended by the NAEP grantee and its subcontractors are devoted to the development of the assessment materials and survey forms, sample design and selection, gaining access to school systems and samples of students, collecting and processing the assessment data, and general project management. About 90 percent of the effort in a typical NELS cycle is expended on these same tasks.

The difference in proportions is due to the inclusion in the NAEP grant of substantial funds for data analysis, reporting and dissemination, technical assistance, and support of the Assessment Policy Committee--activities that either do not occur in NELS studies, or occur at a much lower level. Although NAEP student samples are much larger than those for NELS, the inclusion of supplementary samples make the total number of observations in NELS roughly comparable to the total for NAEP. While the design and procedures for data collection in recent NAEP cycles have been considerably more complex than those for NELS, NELS:88 will collect and process between two and three times the amount of data per respondent than does NAEP.

Despite the differences in allocation, both studies spend a great deal of their resources on similar activities that, if coordinated and integrated, might release funds for improvements in methods and measurement tools, as well as in the quantity and quality of the data they collect. Relatively low levels of coordination, such as establishing a common population definition for the student cohorts studied, will not in themselves result in any economies or have profound effects upon the overall research value of either study. However, it is difficult to imagine the path to much higher and more cost-effective levels of integration --including the extreme of collecting assessment and longitudinal data from the same students--until these lower level issues are resolved in favor of consistency between the programs.

3.2 Population Definitions and Sample Design

Student Samples

At present, NAEP and NELS collect data from different student populations. Historically, NAEP has drawn nationally representative samples students in three age cohorts, and has recently extended the sampling to cover all members of the modal

grade in which the age cohorts were enrolled. NELS studies have drawn samples of students enrolled in specified school grades (not those sampled by NAEP) without regard for their ages. This fact precludes direct comparison of findings between the studies, and limits the extent to which the studies can productively share measurement strategies.

Although many other justifications have been cited, probably the main reason that the original NAEP design chose to sample age rather than school grade cohorts was that it helped to allay the fears of State and local educators that grade-based samples might be used to make inappropriate comparisons across schools, systems or States. Age cohorts appeared more suitable for aggregate analyses because, unlike grade cohorts, they could be objectively and universally defined--their defining characteristics would not vary either over time or across States or school districts. During the nearly two decades of NAEP's operation, Federal, State, and local educators began to rethink the value of this design feature, and in recent assessments called for a combination of age and grade-level sampling. Since NELS studies have always sampled grade cohorts, this approach offered the prospect of comparability in the population estimates made by the two programs, and set researchers to thinking about the possibilities for some form of equating the test scale scores that might be generated by NAEP and NELS.

Unfortunately, however, since NAEP's grade-level sampling was restricted to the "modal grade" for the age cohorts selected, the comparability of the grade cohorts included in the two studies is affected by the way NAEP defines its age cohorts. While the method for defining age cohorts in NAEP has varied between ages (because of the different timings of the assessments for the three age groups) and across NAEP's history, the current age cohort definitions for 9-, 13-, and 17-year-olds result in modal grades of 3, 7, and 11 for the most recent assessment cycle. These differ from the even numbered grades 8, 10, and 12 typically used in NELS surveys.

The majority of educators and policy analysts now agree that the inclusion of grade-level sampling in NAEP will promote secondary uses of the NAEP data for understanding the impact of schools on learning. In many locations, State and district level comparisons using common NAEP scale scores are wanted rather than feared. The question remains which grades to sample. From the point of view of historical continuity, placing NAEP and NELS on a common set of population definitions will require disrupting the time-series of one or the other. Each has its own justification for its current practice. However, NAEP has included grade sampling for only a few assessment cycles, and its approach to determining modal grade has been unstable over that period, whereas the NELS approach to defining its populations and conducting its sampling operations has been constant since 1971, even as its studies have covered wider ranges of elementary and secondary schooling. Defining at least a portion of the NAEP

student populations to be consistent with the NELS 8th and 12th grade cohorts is an essential step for increased articulation between the studies.

A major issue for NAEP designers to face is the desirability of continuing age cohort sampling in light of the current atmosphere favoring greater use of NAEP for comparative analyses. The assessment programs established by most States and large districts focus on grades and the performance standards established for them. If the value of time-series data for age cohorts is considered too great to sacrifice, the age cohort definitions might be revised to produce a chronologically older age cohort. (The cohort of students who turn 13 at any time during calendar year 1986 is older than the cohort of students who turn 13 during the 1986-87 academic year and following summer, and are more likely to be 8th graders in 1986-87.) This choice would result in an odd mix of consistent and inconsistent treatments for each of the three NAEP age cohorts.

Alternatively, NAEP grade cohorts might be defined without reference to the modal grade of the age cohort. NAEP might include 8th and 12th graders in its samples, even if most of its 13- and 17-year-olds (as defined) were in the 7th and 11th grades. However, this would probably result in problems of school-based administration because of the increased mix of students from different grades. Furthermore, under this arrangement, many more students within each sampled school would be eligible for selection because the degree of overlap between the age and grade populations as defined would be significantly lower than under the current NAEP design. If total NAEP sample sizes were to remain constant, the number of 13-year-olds selected would have to be reduced to accommodate the selection of 8th graders, and the number of cases available for 8th grade cohort analysis would be limited by the sampling of younger 13-year-olds. If age sampling must be retained, it will be essential to shift the reference dates for defining age cohorts to increase the overlap with an 8th grade cohort in a given school year.

One of the fundamental requirements for fully merging an assessment program with a longitudinal program is the establishment of a common population definition so that the assessment exercises and longitudinal data on schooling and other experiences may be collected from a single sample of students. If this is to be accomplished in the foreseeable future, either NAEP designers should initiate the process of shifting the NAEP population to the 4th, 8th, and 12th grade level cohorts, and should evaluate the dropping of age cohort sampling entirely, or NELS designers should plan a shift in grade cohort sampling to the 7th, 9th, and 11th grades.

Although the historical continuity of the NAEP time-series would be damaged, in operational terms, the shift to grade level sampling would have little impact on the conduct of NAEP data collection. On the other hand, a shift to odd numbered grades

(7, 9, and 11) is likely to raise significantly the complexity and costs of conducting longitudinal research. Under the current NELS:88 arrangement, the vast majority of 8th grade students will change schools as they move from either elementary, middle or junior high schools to 10th grades in senior high schools. The HS&B experience showed that aside from individual transfers, relatively few students change schools between the 10th and 12th grades. A longitudinal study of a 7th grade cohort would have to cope with a more complex pattern of school changes, which includes a large segment moving from elementary or middle schools in the 7th grade to senior high school for the 9th and 11th grades, and another substantial segment remaining in junior high school for the 7th and 9th grades, and moving to senior high school for the 11th grade. Although (aside from transfers) any given student would not be likely to change schools more than once between the 7th and 11th grades, tracking costs would nevertheless be spread across two followup surveys rather than only one. If longitudinal studies are expanded to include lower grades for the study of early growth, a 4th grade, rather than a 3rd grade start would be more consistent with the traditional two-year interval for followups.

A fully integrated study that collected assessment and longitudinal data from the same students might be restricted to grade level samples, or with suitable revisions in the reference dates defining age cohorts, might retain sampling by both age and grade. However, given its historical focus on school grades as the basic unit of academic instruction, and on the impact of school and classroom effects on the growth of students between grades, an age cohort sample would be of little research value to the longitudinal study. Its presence would tend to complicate NELS questionnaire design and would tend to reduce the size of the sample suitable for estimates of growth between grades. If age based sampling were retained for cross-sectional and time-series assessment purposes, longitudinal followups could be restricted to members of the base year grade cohorts only.

School Dropouts

Both NAEP and NELS designs have included samples of school age youth not currently enrolled in school. NAEP has included broad studies of adult literacy as well as assessments of out-of-school 17-year-olds. NELS began its effort to study school dropouts by sampling 10th graders in 1980 and retaining for followup all sample members who did not remain in school after the base year, and will include a sample of the full population of school dropouts in NELS:88.

To obtain its samples of out-of-school 17-year-olds, NAEP must either draw a sample of households and screen the members for eligible cases, or obtain an eligible sample from another survey with an appropriate design. As learned in the most recent attempt, the development of a sample by screening households is difficult and enormously expensive. Unfortunately, few other

surveys are likely to produce sufficient numbers of out-of-school 17-year-olds for a national assessment. The complex sample of about 30,000 1980 10th graders yielded nearly 2,900 school dropouts by 1982. The NELS:88 sample of 26,000 8th graders may yield nearly 5,000 dropouts by 1992. Since the NELS:88 sample will be "freshened" in 1990 to ensure comparability with the 1980 10th grade cohort, this supplementary sample may contribute additional school leavers.

Although this sample size and schedule may be suitable for assessment purposes, the 1992 NELS dropout population will be limited to members of the 1988 8th grade and 1990 10th grade cohorts who did not remain in school. This population is clearly different from the population of all 17-year-olds who are not enrolled in school in 1992. However, if near future assessments adopt grade-level sampling consistent with NELS, the sampling of school leavers on the basis of their membership in specific grade cohorts may be less problematic. At the least, the NELS sample of school leavers will be fully comparable with the samples of currently enrolled students in that they had progressed to the 8th or 10th grades with their cohorts. Even if NAEP draws an independent sample of currently enrolled 12th graders in 1992, the NELS dropout sample would be comparable to the subset of NAEP 12th graders who had been 10th graders two years earlier or 8th graders four years earlier.

In a fully integrated longitudinal--assessment design, results would be similar to the partially integrated approach. The assessment of out-of-school youth would be an extension of the grade sampling used for currently enrolled students, and would include individuals of varying ages who were surveyed as school attenders during earlier waves.

School and District Sampling

Because of the technical difficulties and enormous expense associated with sampling student populations directly, both NAEP and NELS use multi-stage sample designs and select their students from within nationally representative samples of schools. In the 1987-88, 1989-90 and 1991-92 school years, both studies will be operating in large numbers of public and private elementary and secondary schools. Although neither study samples districts explicitly, each public school selected is connected to a district whose approval to conduct research in the schools must be obtained, just as if the district was a sampled unit. NAEP and NELS frequently operate in many of the same districts, even if they select only a few schools in common. However, at present there is no plan for formal coordination between the studies for school sampling. In part, this is due to the procedures for protecting the confidentiality of research subjects that both programs observe. Moreover, since the two studies sample only small fractions of the population of school buildings, the need for coordination to avoid operational problems has not appeared to be very pressing.

Even if both programs continue largely with their current designs, district level overlap and the move by NELS to begin with lower grade cohorts will make some coordination of school sampling advisable. The adoption of steps for improving the integration between the studies, for example, by establishing a common population definition for students, will almost certainly require more formal coordination of school sampling. For example, if both NAEP and NELS draw multi-stage samples of 8th grade students in the same year, and if both sample designs select schools containing 8th graders with probability proportional to 8th grade enrollment, the chance that schools will be selected for both studies will be greatly increased. Many of the largest urban public school districts already fall into both samples with certainty. The superintendents of these districts contend that their schools are overburdened with requests for access for research purposes--even though in the past the schools selected for the two studies within the districts were usually different. They would find the situation intolerable if the same schools and students were to be selected.

One opportunity for coordinating overlapping samples, the use of rotation sampling of schools, is discussed in at least one other paper prepared for this conference. Under this approach, schools would be selected for multiple years of assessment or longitudinal data collection. After the rotation pattern was established, schools would remain in the sample for six years. Sampling would be phased so that one-third of the total sample of schools would complete its full term of participation in each cycle. At the least, rotation sampling would increase the predictability of the sampling of schools across survey cycles. Taken to the extreme, it could provide a method for sampling schools and students for both assessment and longitudinal purposes, and for regulating the distribution of survey and testing burden across the population.

General Data Collection Strategies

As indicated above, both NAEP and NELS use multi-stage probability sample designs. That is, students are not sampled directly but are randomly chosen from the rosters of enrolled students in nationally representative samples of schools. For the sake of sample design efficiency, the number of clusters (schools) selected in the two studies is large--typically between 700 and 900 schools for each NAEP cohort and about 1,000 schools for NELS. Apart from any oversampling included in the sample designs, this approach results in acceptably high efficiency for student samples.

However, both studies must spend a great deal of time and effort on the process of gaining the voluntary cooperation and assistance of state education agencies and large numbers of districts and schools in which their student samples are enrolled. Once cooperation is obtained, teams of survey administrators must travel to the school sites to collect

questionnaire and test data from students in group sessions. Cost data from NAEP and NELS indicate that approximately 60 percent of the total student data collection resources for each study are consumed by the process of setting up school-based administrations, with only about 40 percent dedicated specifically to collecting and processing student responses.

Although there is considerable overlap between the studies of districts drawn into the samples, the number of overlapping schools is quite small. In a year when both NAEP and NELS are in the field, there may be as many as 2,700 separate schools selected for both studies, consuming between $2 and $3 million in access and setup costs before the first student is surveyed. If the organization of data collection for the two studies were combined--even if different types of data were collected for each--access costs could be substantially reduced, freeing resources for more pressing research problems.

Sampling efficiency considerations dictate the use of many clusters and small cluster sizes. However, the strategy of collecting data at the same sites where samples are drawn builds in operational inefficiency. When both studies are in the field, up to 2,700 separate survey sessions must be scheduled. Survey teams must spend time and resources travelling to each site, and many sessions must be scheduled on the same day, increasing the number of teams to be hired, trained and supervised.

Although the small group administrations are more efficient than individual interviews or test sessions with students in their homes, efficiency could be improved if students from many sampling points could be surveyed and tested at the same sites. Using fewer centrally-located data collection sites would reduce the number of teams and supervisors needed, cut travel and setup costs, and reduce the burden on crowded schools. It would also tend to standardize the survey and testing experiences of the students across schools, and, if handled properly, may serve to boost school and student cooperation rates in both studies.

Among the most important benefits to central site administration is that it substantially raises the feasibility of changing the data collection technology away from printed questionnaire and test booklets, filled out with soft lead pencils and converted to machine readable form by optical scanning or intelligent data entry, to computer assisted survey and test administration. Introducing automated systems within the current data collection approach would mean carting the hardware and technically competent staff to as many as 2,700 sites over the period of a few months, requiring sufficient devices and personnel to staff multiple concurrent sessions at different locations. A smaller number of central sites (at perhaps 200 to 300 locations) could be equipped to handle much larger numbers of students with a much smaller quantity of equipment and many fewer technically expert staff.

Obviously, this approach would work best in urban areas with high density of sampled students. Survey and testing centers might conceivably be established inside the largest, best equipped schools, at district-level facilities, or at independent sites. In rural areas, advantage could be taken of facilities already set up to serve geographically dispersed students, such as area vocational schools. The safety and attractiveness of the surroundings, and the easy connections to existing transportation facilities are the main requirements in either urban, suburban or rural settings.

The establishment of centralized survey and testing sites will help to solve a problem now specific to NELS studies as they choose to begin with lower grade cohorts. In NELS:88, nearly all 8th graders will change schools between the 8th and 10th grades. This transition will tend to generate many more potential data collection sites (schools) and much smaller cluster sizes than in the base year survey. Conceivably, students who attended 1,000 schools in 1988 may disperse into 2,000, 3,000, or even more schools by 1990. If the next longitudinal study design calls for surveying 6th grade, or even 4th grade students, the number of schools generated by the time the students reach secondary schools could become extremely large. This problem is usually handled by disproportionate subsampling of students who do not change schools in relatively intact groups, reducing the efficiency of the design. However, since most students who change schools will do so within a reasonably compact geographical area, the existence of survey and testing centers may obviate the need for subsampling.

CES staff has indicated interest in trials of this approach as early as the field test for the NELS First Followup in 1990. At present, however, planning appears to be restricted to trials of central locations, and will not include tests of automated techniques. If this method succeeds, consideration should be given to establishing sites large enough to handle NAEP assessment activities as well as NELS sessions. Even if the school samples for NAEP and NELS cannot be designed to overlap in acceptable ways, the reduction in the total number of survey sites will dramatically reduce the complexity and cost of data collection for both studies.

I believe that the opportunities presented by survey and testing centers will not be fully realized without the adoption of automated data collection methods, and I urge the Education Department to consider initiating field trials of this approach as soon as possible. If the use of central sites equipped with computer-based survey and testing technology proves feasible, it may open the way for the truly dramatic way to articulate the two studies described next.

Individual Data Collection and Testing and Assessment Strategies

Apart from differences stemming from the longitudinal versus cross-sectional orientations of NELS and NAEP, the most fundamental difference between the two programs is in the basic approach to collecting data from sampled students and in the resulting data structure. NELS is a conventional, individual-level survey design. Except for the differences created by skip or routing patterns in the NELS questionnaires, identical survey and test data is sought from each sampled student in a given survey wave. When questionnaire or test material is changed in followup surveys, the new material is presented in a single standardized survey or test form to all students. Of course, the multi-stage sample design permits the aggregation of student responses to support school level estimates. Nevertheless, the most common analytical unit for cross-sectional or longitudinal research is by far the individual student.

Because each student is expected to answer all questionnaire and test items developed for any wave of the study, the total number of possible items must be kept small. While this causes some problems for questionnaire development (the number of topics recommended by advisory bodies is always several times larger than the available space), cognitive test development is much more difficult. For any student, the small set of items for each test subject may be much too easy or much too hard, resulting in ceiling and floor effects as well as other inefficiencies. The task of gauging individual growth between the 8th and 12th grades with 15 to 20 minute fixed tests is one of the most difficult challenges in NELS:88.

NAEP's testing strategy is radically different. Since NAEP was designed primarily to yield descriptive statistics on the percentages of students who can and cannot answer specific subject matter questions, and since very little inter-item or explanatory analysis was envisioned, there was no need to expose every student in the NAEP sample to all the assessment exercises or background questionnaire items. Conditioned by the political environment at the time of its founding, and by the charge to obtain response distributions for large populations of assessment items in preference to a single set of measures across all students, NAEP first used matrix sampling and more recently a balanced incomplete block design for assigning different mixes of assessment items to individual students. The pool of items for each assessment area is so large that any one student encounters only a small fraction of the complete pool. Spreading the pool of items across very large student samples keeps response burden for individuals at a very low level.

In the early NAEP cycles, even the construction of scale scores was eschewed to avoid the appearance of supporting school, district, or State level comparisons. However, this extreme position has been declining steadily in importance. Inevitably, educators and policy makers began to demand reports of assessment results by a variety of demographic and behavioral

classifications, and encouraged the expansion in the amount of background data collected. However, to avoid significant increases in student response burden, like the assignment of assessment exercises, most background items were "matrixed" or "spiralled" across the sample, and few background items were asked of every student.

NAEP's assessment and survey strategy continues to evolve. The 1986 assessment moved away from the more extreme use of BIB spiralling used in the 1984 assessment and toward a simpler, more manageable design. Nevertheless, NAEP is still a long way from producing rectangular cross-sectional datasets. NAEP's critics continue to complain that the lack of data on individual and school characteristics and the extreme emphasis on measurement of performance on large numbers of curricular objectives cripples the broader utility of NAEP for more powerful policy analyses. The analytical returns, they argue, are too meager in comparison to the resources consumed by the program.

Improving the situation significantly demands three types of change in the NAEP design, all of which would open channels for articulation with NELS. First, NAEP should include the collection of much more contextual data on the schools and classrooms in which its sampled students are enrolled, on the academic performance of the students over time, and on the educational support systems in students' homes. None of these objectives would add further to the burden upon students, but would increase the time taken from parents, teachers, and school officials. Student performance records should be obtained from cumulative files in the form of individual report cards or school transcripts.

Second, to increase its utility for explanatory analyses, NAEP should take much greater advantage of its data collection opportunities by increasing and eventually doubling the response burden it places upon sampled students. Considering the very large sunk costs associated with gaining access to the schools and organizing the testing sessions, NAEP extracts too little data on either learning contexts or students' characteristics and behaviors to justify the effort on policy research grounds. Political leaders may judge NAEP's "national report card" function to be worth the price. However, very few applied education researchers have made significant secondary use of NAEP data to advance educational improvement. Much more student level background and assessment data should be collected.

Third, NAEP should move much farther than it has to date toward a standard cross-sectional survey design to collect questionnaire and test data from students. All additional student background data obtained by increasing response burden should be in fixed format--with the same items collected for all students. If the size of the assessment component is expanded for each student, a significant portion of the assessment exercises should be assigned to all students.

Ideally, this component of the assessment should be large enough to establish crosswalks between NAEP and other tests, specifically those for NELS studies.

I cannot automatically assume, however, that the current NELS test will prove effective enough to merit crosswalks with NAEP or other test batteries. Since NELS:88 is still in its development stage, it is not yet clear that the fixed-format, item-overlap test design being created for 8th, 10th, and 12th grade testing will succeed in measuring student growth in the four subject areas. The main risk is that the limited testing time available will not permit use of enough items in each subject area to measure students' developed ability at each grade with sufficient precision. Since real gains across the NELS grade spans are likely to be of modest size, imprecise measurement may render the test scores useless for longitudinal analyses of school effects upon learning. With a 75-minute test, the NELS testing burden is already considerably larger than that for NAEP (at 45 to 50 minutes). Since the NELS questionnaire requires an additional hour of a student's time, the option of increasing testing time is not feasible.

The use of adaptive testing methods, in which items of appropriate difficulty are presented to students based on their measured ability levels, offers a promising solution. Some forms of adaptive testing can be done with printed booklets and conventional group administration, for example, two-stage testing in which a student's performance on a brief, first-stage "locator test" determines the difficulty level of the second-stage test booklet assigned to the student. However, two-stage tests are less adaptive, and therefor less precise than fully adaptive methods, in which the difficulty of each subsequent test item presented to a student is determined by the students' ability to correctly answer all prior items. This form of test can only be administered on a large scale with computer assistance.

Most of the techniques required for computer assisted survey and test administration have been available for several years. They have not been applied in large scale studies primarily because of the formidable logistical problems of bringing the machinery and the survey respondents together. Simply imagining the enormous complexity of scheduling the movement, erection, and removal of large numbers of survey and testing stations in thousands of schools during a period of a few months has thoroughly inhibited even the initiation of small scale field trials. However, the establishment of a system of fully computerized central sites for use by NAEP, NELS and other Federal, State and local studies makes the problem much more manageable. With central sites, the task of bringing research subjects to the research facility is no more difficult than transporting them to a sports or artistic event--or even to their own schools.

Computerized survey and test administration offers much more than the ability to use adaptive testing. Automated survey

administration would virtually eliminate common response errors in self-administered surveys such as edit failures and missed skip patterns. With minimal keyboard skills, young students could enter open-ended responses for such items as personal background and locating data. The survey and testing software could be linked to various types of audio, video, or input devices that would permit the studies to collect data from certain types of handicapped students. Processing of the student questionnaire and test data, including all data conversion, coding and cleaning, would be instantaneous. Analysis tapes could be created within a few days after the last student logged off his station.

NAEP testing, to the extent that it continues to rely upon distributing pools of items across its student samples, will be free from the restriction of ordering its item groupings into printed booklets. The opportunities for control over complex testing designs may spur the development of entirely new assessment designs, possibly including the use of adaptive testing for efficiently establishing unidimensional performance scale scores for each student. Such scales might be linked to some or all scales established for NELS studies. Driven by the proper algorithms, computer assisted testing may offer enough measurement efficiency to permit the collection of both traditional NAEP and NELS test data in little more testing time now taken by NELS alone. If the feasibility of this approach can be demonstrated, the last obstacle to a fully merged study may be removed, and with it, the elimination of vast duplication of effort.

4. Fully Merging NAEP and NELS

Provided consistent grade-based population definitions and a mutually acceptable assessment and test design can be established, a fully merged longitudinal--assessment design could be initiated in the spring of any even numbered year in the 1990's. Base year studies for three parallel longitudinal studies of 4th, 8th, and 12th grade cohorts could be conducted in approximately 1,000 public and private schools for each cohort. The 3,000 schools of a merged sample is only slightly larger than the expected number of about 2,700 schools to be sampled by the two programs in 1988. By 1992, the two separate studies may be operating in over 3,500 schools. To control the costs of gaining access to schools, to improve the efficiency of recapturing sampled students in followup waves, and to enhance the study of school choice behavior, it would be desirable to select the three levels of school from the same or adjacent public school districts or private school groupings. Once the sample was selected, there would be little difference in the procedures for gaining the cooperation of States, districts, and schools from those now used by either program.

NAEP typically samples over 30,000 students per cohort; NELS studies have varied between 23,000 and 35,000. Sample sizes of about 32,000 per cohort (including any samples of linguistic or

ethnic minorities or other policy relevant groups) should provide adequate precision for all typical NAEP and NELS analyses.

If computer equipped central sites are in place, and highly efficient automated testing procedures have been developed, it should be possible to collect NELS-style questionnaire data and both NAEP-style time-series assessment data and NELS-style change-oriented test scores in a single 3-hour session with each sampled student during the period between February and June of that year. Under any other scenario, the problems of integration quickly multiply. If computerized sites are not available, and school-based group sessions using printed survey forms must be held in schools, it would be impossible to collect NELS and NAEP test data in the same sessions. One option would be to construct a single "optimized" test that would attempt to serve both functions in a 90-minute testing session. A second option would be to retain two types of test, but to attempt to organize them in a manner that would permit the measurement of gains within a school year, by holding a pre-test session either in the fall of the same academic year or in the spring of the preceding year.

In light of current experience, neither of these options seems feasible. Brief paper and pencil tests can barely cope with the research demands of each study alone. A third option would therefore be to develop an entirely new paper and pencil test that would abandon the past approaches to assessment and longitudinal measurement of growth, and attempt to serve both purposes with a new framework. The so called "duplex test design" developed by Bock, Mislevy, and their colleagues, and described more fully in another paper prepared for this conference. While clearly more feasible than the other alternatives, and while sharing some of the item sampling techniques of the traditional NAEP approach, this option exacts the price of abandoning historical continuity with the results of past assessments and longitudinal studies. In return, however, it offers a means for evaluating the progress of schools as well as students toward meeting established curriculum objectives. Running perfectly counter to NAEP's earliest orientation, the duplex design would begin to provide assessment data useful for improving schools that so many educators, administrators, and researchers have been demanding.

Configured as a two-stage test, the duplex design can be (and has been) administered using printed booklets and standard group testing procedures. However, as an adaptive design, its efficiency could be substantially increased through computerization. If the properties and resulting scores from the duplex test design are preferred to traditional NAEP and NELS test designs, use of this testing approach would be perfectly compatible with computerized data collection at survey and testing centers.

Even if the elements of base year design for a fully merged study are finally settled, many challenging complications arise for the next assessment or longitudinal followup (Wave-2), which,

in the normal NELS progression, would occur two years later. Most members of the three grade cohorts would have progressed to the 6th and 10th grades, or would be two years beyond high school; no previously selected students would be in grades typically assessed in NAEP. New samples could be drawn from the 4th, 8th, and 12th grades, but without offsetting reductions, this would double the number of students being surveyed at Wave 2, greatly raising costs in comparison to the current separate designs. One or more of the longitudinal studies cohorts might be dropped, or all three reduced in size through subsampling, but this would defeat many of the purposes of the design change for later followups.

There are two radical but less costly alternatives, both serious departures from current practice. First, the interval between both assessment cycles and longitudinal followups could be shifted from two years to four years. This would allow most of the members of the initial cohorts to advance four grades to the next grade usually assessed in NAEP. It would also leave an unacceptably long spell between assessments and followups. Student tracking problems are difficult enough over a two-year interval, the extension to four years would inevitably increase the attrition in the original sample, probably requiring the selection of additional replicates of Wave-2 students in each cohort. Also, the four year period over which students would be asked to report behaviors is simply too long for acceptably accurate recall.

Second, NAEP assessments at Wave-2 could be carried out on the longitudinal cohort members who advance to the 6th and 10th grades, on those retained in grade, on dropouts and others who left school, adding a new dimension to the scope of NAEP products and reports. This alternative avoids drawing additional samples but imposes the moderate burden of developing tests for grades and ages not now assessed. It is preferable to the first alternative in that it returns assessment data on a two-year cycle, but does so for non-typical grades at alternating cycles. This alternative imposes no unusual change on the operation of the longitudinal component of the study.

Two years later (Wave-3), most of the original 4th and 8th grade cohorts will have reached the 8th and 12th grades and may be reassessed in "normal" grades. However, these Wave-3 samples will represent only 8th and 12 graders who were in 4th and 8th grades at Wave-1, not the full populations of 8th and 12th graders of that year (which will include newly arrived immigrants as well as students not promoted on schedule. A similar problem would have existed at Wave-2, and is already being faced by NELS:88. Procedures are being designed to augment the 8th grade cohort of 1988 with a supplementary sample of 10th graders in 1990 in order to insure comparability with the HS&B 10th grade cohort of 1980.

Across all followup cycles, population mobility and other causes of school change by students, present major operational

problems, reducing the number of originally sampled students who may be assessed and resurveyed in subsequent years. In the absence of empirical data, these problems are easy to exaggerate. The problem of school change between the 10th and 12th grades was relatively minor for the first followup of HS&B, but this time span was short and covered a grade transition across which school change is uncommon. The problem for NELS:88 will be much greater, and the solutions developed to handle it more relevant to the problems of a merged study. The difficulties are greatest if it is necessary to continue to collect data in school-based sessions. Student mobility can reduce the number of students attending the same school to a level too small to justify collecting data at the school. The use of central sites for data collection can greatly reduce the impact of the problem by bringing sampled students from any schools in a reasonably large geographical area into the testing center.

The effects of student mobility may also be mitigated by selecting all three grade cohorts from schools in the same or adjacent districts, or the same general location. The proximity of the selected schools will increase the likelihood that students who change schools at grade promotion will move to a school already in the sample.

By the time of the third wave it will be necessary to draw a new sample of 1,000 4th grade schools (and districts) to restart the cycle. These selections should be made from different geographical areas in order to rotate the survey burden across all eligible units. The total number of schools in the sample would thus remain at approximately 3,000, but the number of geographical areas would increase somewhat, depending upon the specific sites selected.

The total number of students surveyed at Wave-3 would also increase. The size of the increase would depend upon the degree of subsampling for the eldest cohort (12th graders at Wave-1) after leaving high school (Wave-2). The generally higher per-case costs of post-high school survey administration has resulted in reduction of HS&B cohorts to 12,000 to 15,000 (less than half their initial size) after the 12th grade survey wave.

5. Conclusions

The Implications of a Merged Longitudinal--Assessment Design for Education Policy Analyses

Throughout NAEP, the assessment of educational performance has been limited to time-series analysis of national and regional (and, increasingly, State level) descriptive statistics. While this approach may be adequate for gauging aggregate change, apart from relating measured changes to shifts in the composition of the population (such as changes in immigration levels, higher or lower dropout rates, etc.) it contributes little to our ability to understand the impact of changes in educational processes upon

changes in assessment outcomes. Moving NAEP to an individual-level survey design would provide a partial solution by enriching the kinds of time-series analyses that might be performed.

However, efforts to reform educational practice increasingly rely upon intervention programs that are designed to affect students for a period of years, not simply at one point in time. Understanding how intervention programs work on exposed students is thus a longitudinal research issue. If most intervention programs were designed to serve narrow purposes for special populations, they would be better understood through the use of highly focused, stand-alone studies. During the 1980's, however, intervention programs have become very broad in scope and tend to focus upon the same basic elements of educational progress that historically have been at the core of NAEP. By combining the features of assessment design with longitudinal survey methods, a merged study would not only provide for all of the usual cross-sectional and time-series measures and analyses, but would also extend the analytical power of the database to explore the dynamics of change in the performance of individual students.

To some extent, the NELS program already collects much of the data necessary to support analyses of this type. Unfortunately, despite its large size and cost, it has not yet had sufficient resources to collect sufficiently accurate or precise estimates of students' academic abilities and performance over time. At most, NELS tests have provided valuable classifications of gross achievement levels, and for some subject areas, broad measures of stability or change in scores. For the full potential of the longitudinal design to be met, we must find the means to do a better job of testing sampled students.

By combining NELS design strengths with a testing program of a quality suitable for assessment reporting, we would be much closer to the goal of understanding what works in educational reform. Although I have not treated the possibilities in detail here, I believe that a merged design should also preserve other features of the most recently designed longitudinal study, namely the collection of data from parents and teachers, and school principals of all sampled students, and should include the collection and processing of report cards, transcripts and other student records. The data structures provided by this design would come very close to satisfying the long range goals and requirements of a truly integrated system for studying elementary and secondary education. However, these ambitions may never be fulfilled unless the first small steps toward articulating the longitudinal and assessment programs are taken in the very near future.

Instrument Design for a Combined NAEP and NELS

R. Darrell Bock
The University of Chicago

That the National Assessment of Educational Progress (NAEP) and the National Educational Longitudinal Study (NELS) are both continuing, large-scale surveys of current educational outcomes in the United States is perhaps reason enough for exploring the possible advantages of combining them into a single data gathering effort. The task will not be without obstacles, for objectives of the two surveys as originally conceived are quite different. Whether a combined NAEP and NELS, collecting one body of data, could serve the manifest purposes of each remains to be determined.

NAEP was designed primarily to monitor long-term trends in attainment in a wide range of subject matter from the primary, middle, and secondary school curricula. Skills and topics considered to be vital to the national interest--reading, science, mathematics--have been repeatedly surveyed, supplemented occasionally by assessments of literature and the arts. In its original conception, NAEP reflected the concern for curriculum and instruction of its founder, Ralph Tyler, who saw the survey in two main roles: on the one hand, as a tool for high-level policy and planning in education, and on the other, as a source of data for subject-matter experts, curriculum specialists, instructional planners, textbook writers, test constructors, and students of education generally.

When a national assessment based on this conception began to function, it became apparent that these two roles demanded rather different forms of data. Persons responsible for policy decisions required only very general measures of educational outcomes. They wanted to know whether the schools were producing a sufficient number of trained and educated young people to meet the needs of a developed society. They viewed educational assessment as a source of census-like information on the educational attainment of the citizenry.

Specialists in curriculum and instruction, on the other hand, needed much more detailed information about educational outcomes. They wanted to know the suitability of particular subject-matter at given grade levels, the order in which new topics should be introduced, and the time and resources that should be devoted to each. Similarly, exercise writers and test constructors needed statistics on item difficulty in order to locate their materials at appropriate points in the instructional sequence. NAEP was the first attempt to provide such information routinely at the national level.

In contrast to NAEP, which has surveyed temporal change in the cross-sectional measures of society-wide educational

performance, the NELS survey has concentrated on obtaining longitudinal records of the development of individual learners as they pass through the school system and enter adult life. School attainment, as measured by formal achievement tests, was viewed an important but not exclusive aspect of this development. Other facets of attaining social and economic maturity and independence have been included in the survey. The NELS approach is necessarily much more oriented toward "trait" measurement than is assessment and is not intended to evaluate attainment of detailed curricular objectives. Whereas NAEP can produce data on many subject-matter topics in greater or less detail, the limitation of testing time in NELS allows only for overall measurement of attainment in four broad subject matters--reading, mathematics, science and social studies. Although the school programs and courses pursued by the learners have been recorded, there has been no provision for measuring curricular objectives at the level of detail required for the mundane decisions of instructional planning and material preparation. Unlike the NAEP evaluation effort, NELS is a research study aimed at elucidating the causal factors in the individual learner's attainment, persistence, school participation, and emotional and occupational success. The NELS design is longitudinal at the level of the learner and cross-sectional only with respect to those institutional variables that are measured on each occasion of interviewing and testing.

1. Instrument Design for Educational Assessment

The contrasting purposes of the NAEP and NELS surveys are reflected in the design of their respective data gathering instruments. In NAEP, the assessment of attainment of curricular objectives depends upon the use of so-called "matrix-sampling" schemes that permit the evaluation of a large number of topics without excessive demands on the respondent's time. A typical assessment instrument consists of many forms--30 to 40--each containing a few items, or perhaps only one item, from each curricular objective to be assessed. Because students can respond to between about 25 to 45 items in a 40-minute testing session, each test booklet can provide information on a comparable number of objectives. Because the multiple test forms contain different items assigned randomly to students, the measurement of each curricular objective (which can include both cognitive and effective outcomes) can be based on perhaps as many items as there are distinct test booklets. When there are many such booklets, this type of design assures the degree of generalizability required for dependable results at the school, district, state or national level.

To appreciate the importance of the high degree of generalizability at the group level, it is instructive to examine data from the California Assessment Program measuring curricular objectives at the school level. Table 1 shows the correlations between unweighted school means for successive years of attainment in sixth grade reading as a function of the number of

students in the school at grade level and the number of items in the test instrument.

TABLE 1

Effect of sampling of students and sampling of items on the year-to-year correlations of mean sixth grade reading attainment scores of California schools*

		No. of items in matrix sample		
		85	128	400
Median number of students sampled per grade	50	.59	.73	.79
	100	.67	.78	.88
	200	.76	.81	.93

*From Bock & Mislevy (1986)

Although it is well known that the decreasing sampling variance in the means as school size increases will result in a reduction of attenuation of the year-to-year correlation, it is perhaps less appreciated that the increased measurement precision due to larger samples of items has the same effect. Thus, the justification of the matrix-sampling design is its capacity, not only to measure many and detailed curricular objectives, but also to measure with greater precision in broad subject-matter areas by aggregating over large samples of items.

2. Instrument Design for Longitudinal Research

The typical assessment instrument does not, however, provide any well-defined or dependable information about specific knowledge, skills or attitudes of individual students. We cannot expect the designs thus far implemented by NAEP, or by the state assessment programs, to serve the purposes of a longitudinal study such as NELS. For those purposes, the measure of attainment must be more in the nature of a conventional educational achievement test: it must make use of perhaps 40 to 50 items in each of a relatively small number of subject-matter areas in order to obtain dependable scores for purposes of measuring growth and relating achievement to other case-level variables.

Such information can, of course, be aggregated to the school and higher levels to study institutional and community variables with high precision, but relatively little detail concerning curricular objectives can be retained. This is perhaps just as well, since those research specialists who will use NELS data for

process studies or causal modeling are unlikely to pursue curriculum-oriented research in more detail than the measures of overall subject-matter attainment provide. Even to obtain that level of detail with sufficient precision for individual measurement requires considerably more testing time than is required by assessment instruments. The reading section of a recent edition of the Stanford Achievement Tests, for example, contains 45 vocabulary items and 40 reading passages--twice the length of an assessment instrument that would span several subject-matter areas in 40 minutes of testing. Similarly, the present NELS achievement test which covers four subject matter areas, require 70 minutes (two class periods) for administration.

Although it has never been previously suggested for educational research, a strong case can be made for enhancing the generalizability of individual achievement tests by employing multiple, stratified randomly parallel test forms similar to those in matrix sampling. If such forms are accurately equated, scores based on them can be used in statistical analysis without identifying the forms and can increase precision at the group level equal to that demonstrated for matrix-sampled data in Table 1.

These observations suggest that it should be possible to design an instrument that could combine the goals of curricular assessment with those of individual measurement for purposes of longitudinal and process-oriented research. Such a design, based on work of Bock and Mislevy (1986), is proposed below. Before it can be described, however, an important distinction in the measurement of educational attainment and other characteristics of the learners needs to be clarified.

3. The "Fixed" and "Random" Item Concept

In designing instruments for surveys such as NAEP or NELS, it is important to distinguish between items that measure specific characteristics of the respondents and their demographic background, as opposed to those that measure attainment by sampling the skills and content of school subject-matter areas. From a statistical point of view the former are "fixed", and appear simply as variables in a qualitative or quantitative multivariate model, while the latter are "random", and freely exchangeable with other items sampled from the specified content domains. Although the population "percents correct" for a few illustrative attainment items are sometimes discussed in assessment reports, they are meaningful only as representative of the specification class from which they are drawn. The relevant statistic is the mean, or scale score, estimating the central tendency of the class. Such a statistic is, of course, subject to error variation due to the item sampling. It is that variation that attenuates the correlation of attainment outcome measures among themselves or with the demographic variables. Variation arising from the sampling of items can be suppressed either by lengthening the test administered to the individual

respondents or, where time constraints and the large number of content categories to be assessed make this impossible, by resorting to matrix sampling and reporting results only at the group level.

The implication of the fixed- and random-item concept for assessment surveys such as NAEP is that, while multiple forms of the survey instrument must be produced to provide for the matrix sampling of items that measure educational attainment, those parts of the form that measure fixed characteristics of the respondent or the respondent's background should be the same on all forms. In the analysis, each of the latter, fixed items can serve as a distinct explanatory variable, whereas the former, random items appear as outcome variables only when aggregated into statistics representing content categories or curricular objectives. As suggested above, an instrument for surveying attainment at the individual level for a study such as NELS can, nevertheless, benefit from improved dependability of measurement at the group level if the cognitive section of the instrument is produced in multiple, parallel test forms. For longitudinal measurement, the forms must be composed of a series of subtests "vertically" connected so as to measure growth over a range of years. Any given respondent would be tested with a subtest appropriate to his or her years, drawn from the same form. Any two respondents might not, however, be responding to the same form. Collectively, the several forms would contain a sufficient number of items to assure good generalizability and to reduce attenuation due to measurement error in any estimation of relationships at the group level.

4. The Duplex Design

For some time it has been thought that the evaluation of progress in obtaining detailed curricular objectives could not be carried out in the same study with the reliable measurement of individual student attainment. State educational testing programs that require both types of information have always obtained them from separate, independent testing programs. Although the items appearing in the instruments of these distinct programs may be rather similar, the manner in which the items are assigned to forms and scored is quite different.

Recently, however, Bock & Mislevy (1986) have proposed a new type of instrument, called the "duplex design", that serves both the purposes of assessment of curricular objectives and those of measuring individual attainment. The design exploits certain developments in modern testing practice, based on item response theory (IRT), that make possible the scoring of multiple test forms comparably on the same scale and suggest strategies for reducing the number of items required for good precison of measurement at the individual level. Basically, the duplex design is a multiple-form instrument that permits measurement at the classroom or school level for as many distinct curricular objectives as there are items in each test booklet. But the

items are assigned to the booklets in such a way that each booklet covers uniformly all of the content categories. Aggregating responses over suitable item subsets within each booklet produces reliable scores for individual subjects in the more important areas and subareas of the subject matter. At the same time, aggregating responses across the booklets for each item specification class provides detailed information on curricular objectives at the school level.

As an example of a duplex design, the item structure of a test form for an eighth grade mathematics assessment is shown in Table 2. The item classification is an example of a so-called "content x process" specification of curricular objectives similar to that used in the studies conducted by the International Educational Achievement Association. Its rows correspond to content organized under main mathematics topics, and its columns correspond to the cognitive process involved in responding to items representing the curricular elements defined by the row and column intersections. In this particular design, the content categories are, in effect, a union of the eighth grade mathematics curriculum of the states of California and Illinois, together with those presented in current textbooks and curriculum guides in the mathematics education field. The "processes" represented by the columns of the table, referred to here as "proficiencies", reflect long-standing distinctions in the cognitive research literature--namely, 1) procedural knowledge exhibited in productive skills, 2) factual knowledge of terms, relationships, and concepts, and 3) higher-order thinking processes involved in the application of knowledge and procedures to problems of reasoning, proof, and real-world applications. The design, when analyzed by IRT methods, produces scale scores for individual respondents in each of these three proficiencies and five content topics, and, at the group level, scores for each of the 48 content x process elements in the table. The psychometric model by which this analysis is carried out links these two types of scores together so that they appear on the same scale (Bock & Mislevy, 1986). Thus, the group-level mean of the proficiency and content topic scores aggregated over respondents equals the mean for the corresponding curricular objectives aggregated over elements within proficiencies.

5. Two-stage Testing

Another feature of this design, which also exploits modern item response theory, is the provision for adaptive testing by means of two-stage administration. For more precise measurement of proficiencies at the individual respondent level, each form of a test, such as that in Table 2, consists of three test booklets containing, respectively, items at the easy, intermediate, and hard levels of difficulty, respectively.

Before the respondents take the main test, they are administered a short pretest, the results of which determine the booklet of the main, or second-stage test to be assigned to each

examinee. This makes the measurement more reliable by insuring that each examinee is responding to items approriate to his or her level of attainment. Four common items appearing in the easy and intermediate forms, and in the intermediate and hard forms, link together the scales for the three levels of the test booklets.

TABLE 2

A GRADE 8 MATHEMATICS DUPLEX DESIGN

Content Categories	Proficiencies		
	a. Procedural Skills[1]	b. Factual Knowledge[2]	c. Higher Level Thinking[3]
10. **Numbers**			
Integers	11a	11b	11c
Fractions	12a	12b	12c
Percent	13a	13b	13c
Decimals	14a	14b	14c
Irrationals	15a	15b	15c
20. **Algebra**			
Expressions	21a	21b	21c
Equations	22a	22b	22c
Inequalities	23a	23b	23c
Functions	24a	24b	24c
30. **Geometry**			
Figures	31a	31b	31c
Relations & Transformations	32a	32b	32c
Coordinates	33a	33b	33c
40. **Measurement**			
English & metric units	41a	41b	41c
Length, area & volume	42a	42b	42c
Angular measure	43a	43b	43c
Other systems (time, etc.)	44a	44b	44c
50. **Probability & Statistics**			
Probability	51a	51b	51c
Experiments & surveys	52a	52b	52c
Descriptive Statistics	53a	53b	53c

[1] Calculating, rewriting, constructing, estimating, executing algorithms.

[2] Terms, definitions, concepts.

[3] Proof, reasoning, problem solving, real-world applications.

A field trial of a two-stage instrument based on the design in Table 2 is now being carried out in the States of Illinois and California under the auspices of the National Center for Student Testing, Evaluation, and Standards, with the support of the Office of Educational Research and Improvement (OERI) of the U.S. Department of Education. (The study is being conducted by NORC, the University of Chicago). The instrument created for this trial makes use of items contributed by the California and Illinois Assessment Programs and by the International Achievement Association. It covers 45 (rather than 48) content categories, three proficiencies, and five content topics in eight test forms composing easy, intermediate and hard test booklets. A separate pretest consists of 12 highly discriminating mathematics items. The total number of distinct items in the second-stage forms is 888.

This form of instrument is designed to be administered in the classrooms for the relevant subject matter area--in this case, eighth grade mathematics classrooms. One or two days before the main testing, a classroom teacher presents the pretest and the answer sheet for both tests to the students. Each student fills out that part of the answer sheet that provides identification and background information, and responds under timed conditions to the pretest items. The student then inserts the answer sheet in the pretest booklet and returns both to the teacher.

Before the day of the second-stage test, the teacher scores the first-stage test and then inserts the student's answer sheet into a second-stage test booklet with a level of difficulty appropriate to the score on the first-stage test. The cover sheet of the second-stage test is trimmed in such a way that the student's name on the answer sheet is visible; at the time of second-stage testing, it is then easy for the teacher to return the test booklet to the correct student. Because the second-stage tests are packed in rotation when supplied to the teacher, the assignment of forms of each level of difficulty to the students is effectively random. This insures that the scores from different forms of the second-stage tests are drawn from the same population (use is made of this fact in the IRT scaling of the test forms discussed below).

For the duplex design to supply information about individual student achievement as well as aggregate information about the school, the classroom time required to cover the main subject matter areas is greater than that typically devoted to school or program assessment alone. The use of two-stage adaptive testing demands less time, however, than would conventional achievement testing of the same subjects. A reasonable schedule for assessment based on the duplex design might be the following: one class period would be devoted to each of three subject matter areas, namely, 1) mathematics, 2) reading, including passages which test knowledge of science, history and social science, and literature, 3) written expression, including multiple-choice tests of the mechanics of spelling, punctuation, capitalization,

and grammar, and writing of several paragraphs in response to a distinct prompt assigned to each test form. These three class periods, which would not necessarily occur on the same day, would be taken from mathematics, history and social studies, and English classes, respectively. It is assumed that the teacher teaching these subjects would administer the test in the corresponding area.

On an earlier day, another class period would be devoted to the students recording their names and background information on the answer sheet and to the administration of a first-stage test consisting of 15 verbal and 15 quantitative items. There is sufficient space for all of this information, including three second-stage tests, each of 45 to 50 items, on the two sides of a double sided answer sheet. Alternatively, expendable test forms could be used at somewhat greater expense.

6. Analysis of the Duplex Design

The IRT methods required to estimate scale scores at the student level and at the school or other group level already exist for multiple-choice items; they are also in an advanced state of development for essays graded by expert readers using ordinal rating scales designed for this purpose. For scaling such data at the individual student level, rigorous maximum likelihood methods are available for multiple-choice items in the BILOG program of Mislevy & Bock and for ordinal rating categories in the MULTILOG program of David Thissen. For scaling the matrix-sampled data at the school or higher level, corresponding maximum likelihood procedures have been developed by Bock and Mislevy for application in the California Assessment Program. These procedures are at the present time being combined into a computer program, called "DUPLEX", which will perform both types of analyses simultaneously and relate the resulting student level and school level scales. For school level scaling of graded ratings of written essays, a program called "ORDINAL" is being written for purposes of the writing assessments of the California Assessment Program. The procedures for binary-scored items make use of the two and three parameter logistic model for item responses (see Lord & Novick, 1968), and the procedure for graded scores uses a variant of Samejima's cumulative logistic model. Experience with the California Assessment Program and from pilot studies shows that these models are flexible enough to fit a wide variety of item responses without restricting item selection.

In the analysis of the two-stage instrument design, each test form, as represented by its easy, intermediate, and hard test booklet, is scaled separately. Because of the item overlap of the three booklets within each test form, a single IRT scale is obtained on which students may be scored comparably from the lowest to the highest levels of proficiency. The origin and unit of such scales for each form are arbitrary and may be set by specifying the mean and standard deviation for the population in the first year of testing. In the California Assessment Program,

for example, the mean was set at 250 and the standard deviation at 50 in the first year of the IRT-based assessment at each grade level. The scale thus defined is then maintained as items are added and retired so that trends over years can be examined on an absolute basis. Inasmuch as the students responding to each form in the assessment instrument are drawn from the same population, setting the mean and standard deviation of all forms to the same value provides comparability of scores from one form to another. Thus, all of the proficiency estimates from the design, although based on different forms are from different levels of difficulty within forms, are comparable for any purpose of student counseling, program evaluation, or statistical analysis.

Scaling at the school level within curricular elements is performed by a multistage IRT model in which the student population is structured into the three pretest groups within schools. For each pretest group an IRT score is estimated on a scale linked by the same common items that link the proficiency scales, and the score for the school is the weighted mean of the pretest group scale scores. The overall origin and unit of the classroom or school level scores is chosen to be consistent with the arbitrary origin and unit of the proficiency scales. The theory of this type of scaling is discussed in Mislevy and Bock (1987).

7. Properties of the Duplex Design

The methods of collecting and analyzing educational attainment data by means of the duplex design can benefit many parties to public and private education. With minimum demands on classroom time, it supplies a reasonably detailed and accurate profile of each student's accomplishments in the main subject matter areas of the school program. Student scores can be expressed either in standard form or in percentiles of populations to which the student might legitimately be compared. These results can be shared with parents and teachers and thus bring some benefit to the student that justifies his or her participation in the testing. The teacher can also receive this information summarized for the classroom, and for the school, as a basis for discussion with students and parents. At the same time, summary information about classrooms and the school for management purposes can be given in detail for each of the curricular elements in the design. This information can be displayed in a profile of strength and weakness in each area of instruction for classrooms relative to the school average, and for the school relative to the total population average. These results have direct implications for the teachers' allocations of time and effort in instruction during the coming year, and thus justify the efforts of the classroom teacher in administering the test forms and scoring the pretest. The classroom setting, the partication of the teacher, the scoring and reporting for individual students, all contribute to more orderly testing and better motivated performance.

Similarly, the information on the distribution of student proficiency scores in the school, and the profiles of scores on the curricular elements, shown in relation to schools with similar community backgrounds, returns something of value to the school principals and superintendents, and encourages their participation in such studies. A serious problem with sampling assessment studies such as NAEP is that they return no useful information to schools or students. Although NAEP of the past has been able to depend upon the prestige of the program to recruit schools, recruitment may become more difficult as the states continue to develop their own assessment programs and demands for testing time increases. Moreover, if NAEP attempts to expand the sampling of schools sufficiently to report state results, then the prestige of participation would be further diminished by the commonness of the occurrence. In contra-distinction, the NELS study, which has the capability of returning results to students and to schools, does not have to depend upon the prestige of the program alone to encourage cooperation. The same strategy could be adopted by NAEP if the duplex design were used to provide a profile of useful scores at the individual student level.

An advantage of the duplex design for curriculum oriented studies is the content x process organization of the instrument. This schema for specifying curricula objectives is the most widely used, often in far greater detail than even the duplex design can accommodate. Some of this detail is, of course, available for reporting at the item level within the item classes represented by the cells of the duplex design. NAEP has in the past provided such information for the specialized purposes of text book exercise preparation and test construction.

For purposes of the sociological and psychological studies that figure prominently in the use of NELS data, the proficiency scores of the design provides a manageable number of variables for describing the educational attainments of each student in the sample. Due to the highly structured parallelism of items in the multiple test forms, all of the measures will have similar psychometric properties and thus behave uniformly when analyzed by conventional least squares techniques that assume homogeneous measurement error. This design will also provide a more complete profile of the individual student's strengths and weaknesses than is available in the highly general, "trait-like" scores that the present NAEP contractor is developing in the reading and writing areas. The duplex design accommodates the multivariate nature of educational outcomes, as opposed to psychological dimensions of primary ability, and does so in a form that is readily accessible to analysis by well-developed multivariate procedures. When combined with the background data at the student, school and community level, the proficiency scores can be supplied to secondary users in a form suitable for multivariate analysis of variance, analysis of linear structural relationships, and multiple correlation in regression analysis.

8. Comparison with the Present NAEP Design

It is interesting in this connection to compare the properties of duplex design with those of the spiraled balanced incomplete block (BIB) design used by Educational Testing Service, the present NAEP contractor. The BIB spiraled instrument for Reading, for example, is not a two-stage test and does not have enough items per form to permit any more than a rough estimate of a very general reading attainment scale score. Moreover, the number of scaleable items varies from one form to another so that the measurement error is not homogeneous from case to case.

The reason for the choice of balanced incomplete block design with a constant number of items linking all pairs of forms was to make possible the scaling of all the items in one massive link-form IRT analysis using the LOGIST program developed at ETS. The pattern of common items in the BIB design facilitates this type of analysis. Another reason for the choice of this design was that uniform joint-occurrence frequencies between all pairs of items could be calculated for purposes of item factor analyses to establish the dimensionality of the item set. Unfortunately, there are so many items and forms in the design that the numbers of subjects in which joint occurrences can be observed are relatively small, only a few hundreds, even though the total sample size is in the order of 20,000. Because item factor analyses typically require considerably larger sample sizes for estimating joint-occurrence frequencies and tetrachoric correlation coefficients (samples of the order of a 1,000), the item factor analyses have little power to reject the hypotheses of unidimensionality. To remedy this situation, future BIB spiral designs planned by ETS will have fewer items and yield better estimates of the joint-occurrence frequencies.

In contrast, the duplex design relies, not on item linking, but on random assignment of forms to respondents and equating at the population level to produce comparable IRT scales for the multiple forms (equivalent groups equating). Because all forms are identical with respect to the item categories represented, studies of dimensionality can be carried out on separate forms where the joint occurrence frequencies are estimated from complete data. Replication of these analyses over forms then provides for verification of the stability of the factor solutions. Unlike the BIB spiral design, the duplex design with its uniform content structure in the forms is robust to minor failures of the unidimensionality assumption required in IRT scaling of the proficiencies. Content is balanced in such a way that a unidimensional scale remains a stable and meaningful summary of the general achievement factor in the presence of small group factors that may appear among some of the content categories.

Differences also exist in the approach to the analysis of the BIB spiral design by ETS and the analysis that has been developed for the duplex design. On the one hand, ETS has made

use of maximum likelihood estimation of scale scores of the respondents to the reading assessment. Because many of the forms had as few as nine scaleable items for the reading scale, and the instrument was not adaptive, there were many cases of subjects with all right or all wrong responses. In the maximum likelihood estimation used by the LOGIST program, these response patterns do not yield a finite score. To avoid this problem, ETS converted the scale scores into expected true scores on the interval 100 to 400, similar to CAP scores. This only serves, however, to convert the indeterminant scores into 100s or 400s, and can lead to U-shaped distributions of scores in the population. To avoid this problem the multiple imputation technique advocated by Donald Rubin has been applied to reconstruct a more plausible distribution of scores having the same means in demographic subclasses as the original data.

In the analysis of the duplex design, on the other hand, each proficiency score is based on some 15 items of the second-stage test and the overall math score is based on 45 items; moreover, the scale scores are posterior means, given the item response pattern for the pretest and main test and making use of an empirical prior distribution estimated at the highest sampling level (for example, at the national level). Because of the highly homogeneous nature of the duplex scales and the robust nature of posterior means, this method of scoring will produce distributions of scores suitable for conventional statistical analysis at all levels of the data without attribution or imputation.

For analyses at the school, community and higher levels, the duplex design will give scale scores for similar good statistical properties for each of the numerous curricula objectives spanned by the design. The precision of the school level scores is, of course, a function of the number of students per school and the measurement error is therefore relatively small for all but the smallest schools. Analyses of this type of data will typically require weighting by the number of students per school (or strictly speaking, by the posterior variance of each score), but most computer implemented procedures for data of this type provide for such case weighting of scores.

9. Longitudinal Studies

Studies of educational attainment can be longitudinal at two levels, both of which are served by the duplex design. Scales at different grade levels can be linked by common items to provide a uniform scale for measurement of growth of individual students followed over a period of time, as in the NELS study. Because the range of the instrument at grade level is wide due to the use of two-stage testing, the same instrument could be used to study change over two or three grade levels by taking advantage of the existence of numerous multiple forms and arranging that the same student take different levels of the same form on each occasion. This same wide range insures that the duplex scales can easily be

linked over a number of grade levels--perhaps from 4 to 8 or from 8 to 12. Certain scales such as reading and writing probably could be linked over the entire range from 4 to 12. The design is well suited to this type of vertical equating based on IRT methods.

At the school level, assessments studies can be longitudinal for the institution, but cross-sectional with respect to the students. This is the case in the California Assessment Program, for example, where progress of individual schools is charted from year-to-year using indices based on the scale scores for the schools in the various curricular areas. For a national study, the same monitoring of progress at the institutional level could be carried out by the use of rotation sampling of schools. That is, a given school would enlist in the study for perhaps four years, and be tested at the beginning, middle and end of that period. In this way short-term gains and losses could be examined over this period and, by use of random regression models, longer scale trends could be modeled with this form of longitudinal data. Typically, one-third of the schools would turn over at each testing period, thus providing generalizability to the school population in the nation, and allowing for the detection of any possible "Hawthorne" effect due to membership in the school panels.

10. Possible Benefits from a Combined NAEP and NELS

Because, at the respondent level, the NELS sample is longitudinal and the NAEP sample is cross-sectional, the two studies can intersect only when the NELS respondents are selected in the first year from the schools recruited for the NAEP panels, (assuming rotation sampling of schools). In the next contact with the NELS sample, the 8th grade students would be, in most cases, in other schools, and the 11th grade students would have graduated or dropped out. Nevertheless, certain economic and methodological benefits will follow if the two samples are coincident, albeit only initially.

Economic Benefits. Significant savings would be realized by having only one sample of schools to select and to recruit into the two studies. More resources could be devoted to recruitment, therefore, and participation rates could be expected to increase. Especially so, considering the rich array of reports that could be returned to the schools from the two studies. Economies of data collection would also result in the first year. Assuming that all students at grade level are tested on the attainment measures for purposes of the NAEP study, the selected NELS students would only have to complete the more detailed background questionnaire. Presumably, the school questionnaire would be identical for the two studies and avoid duplication at that stage. NAEP would, of course, lose some of this initial benefit as new schools not included in the NELS study rotated into the sample, but would gain it again if a new wave of NELS students were selected.

Other major economic benefits would result from common development of the attainment tests, common school questionnaires, and overlap of parts of the student background questionnaire. The quality of these instruments would also gain from the greater resources available for preparation and field testing.

Benefits of a common instrument design. If a thorough going duplex design covering three or four subject matter areas were adopted for the NAEP survey, then NELS would benefit from the greater breadth, precision and generalizability of the two-stage NAEP instrument. Whereas the present NELS design employs only 20 items in one-stage testing of four main subject-matter areas, the NAEP duplex design would employ 40 to 50 items in a two-stage test of each area. In addition, the NAEP instrument would give three subarea scores within each subject matter (for example, the three mathematics proficiencies measured by the design in Table 2). With two-stage testing, the subarea scores would have a precision comparable to the main subject-matter areas of the NELS instrument, while the NAEP main area scores would have a measurement-error variance about one-third that of the NELS tests. Additional testing time would be required for administration of the NAEP test, but it would be justified by the much broader use that could be made of the more detailed data and by its greater relevance to the schools and the students.

NAEP, in turn, would benefit from the more complete background information that would be available from students sampled for the NELS study, and would have an entirely new source of information in the responses to the parent questionnaires. NAEP could also benefit from information on course work and opportunity to learn, which is more thoroughly covered in NELS than in the NAEP surveys. This information is especially important when NAEP results are broken down by region, type of community, or race, where differences in attainment in certain subject matters may merely reflect different degrees of exposure to the content.

Conceptually, both studies would benefit from the possibility of connecting results through common variables and common respondents when the data are analyzed and reported. Secondary users of the data should find these relationships productive in the more complete modeling and checking of internal consistency that would then be possible. Broader social implications of the NELS findings could be deduced from the statistical relationships between the developmental and the institutional variables; conversely, the policy implications of the NAEP results could be buttressed by the plausible explanation of them at the level of learning and social interaction processes. A less tangible, but real, benefit would be the facilitation of reporting, public discussion, and scholarly study that would result from the two studies using precisely the same outcome measures and many of the same measures of background variables. An opportunity exists here to avoid the proliferation

of uncomparable and inconsistent measures that have been a stumbling block for empirical research in social science for more than fifty years. A combined NAEP and NELS study, using common variables to a considerable extent would be a welcome precedent for future work in the field.

REFERENCES

Lord, F.M. & Novick, M.R. (1968) Statistical Theories of Mental Test Scores. Reading (MA): Addison-Wesley.

Mislevey, R.J. & Bock, R.D. (1987) Comprehensive educational assessment for the states: the duplex design (submitted for publication).

Sampling Problems in Merging a Cross-sectional and a Longitudinal Program

Bruce D. Spencer

Associate Professor of Statistics, Education and Social Policy, and Urban Affairs and Policy Research
Northwestern University
and
Director, Methodology Research Center, NORC

1. Maintaining Representative Cross Sections

The first general problem in merging a longitudinal and a cross-sectional program is to maintain a representative cross-sectional sample at each time point. (A representative sample is a probability sample from the target population with known selection probabilities for the units in the sample.)

Part of this problem is easily handled. Consider, for example, NELS:88. NELS:88 will select a representative sample of 8th grade students and 8th grade schools (i.e., schools containing 8th grade students) in the 1987-88 school year. These students will be followed up in 1990, when most, but not all, are in 10th grade. To obtain a representative sample of 10th graders in 1990 we will need to do two things. First, we will need to exclude those members of the base-year sample who are not in 10th grade in 1990. Second, we will pull an additional sample of 1990 10th grade students who were not eligible for selection in 1988--these include, for example, immigrants, certain 1988 9th grade students who repeated a grade and 1988 8th grade students whose schools were not eligible for a base-year selection but who are part of the target population of 1990 10th graders. The result will be a probability sample of 1990 10th graders.

A variety of sampling strategies exist for modifying a base-year sample in this way. One strategy uses linking techniques to obtain a sample of 1990 10th graders without the necessity of drawing a representative sample of 1990 10th grade schools; for further discussion see Appendix A. If one does not need a representative sample of schools for times other than the base year, that strategy has the virtue of simplicity and ease of execution. If a representative sample of 10th grade schools is needed, then the sample selection of 10th grade schools must be performed in such a way that the school selection probabilities can be calculated and also (for efficiency) are roughly proportional to the size of the 10th grade class. While these two considerations induce complexities into the sample design, a variety of designs can be developed to accommodate them; see Appendix B for some examples. However, these designs are complex and more expensive to use than is the linking design. Current plans are to use the linking design for NELS:88.

This discussion of maintaining representativeness in the cross-sectional samples has been set in the context of a base-year sample of 8th graders followed up two years later, but the same issues would pertain to a later four year follow-up (12th grade) or to other base-year cohorts (such as 4th or 10th grade). To sum up, a variety of techniques for maintaining representative samples of students are available and usable, but maintaining representative samples of schools is costly.

There are, to be sure, some problems inherent with maintaining a longitudinal sample of younger students over time. As a sample of students graduates from elementary schools to middle schools or junior high schools to high schools, the number of school buildings in which they attend school will tend to increase. Thus, if 28,000 students in 1,000 schools composed a 1988 8th grade sample and if all of the sampled students were resurveyed in school in 1990 it might be necessary to visit more than, say, 2,000 or 3,000 schools. That would be unduly expensive, and subsampling strategies will be explored to control field costs and maintain statistical efficiency. Note, however, that this problem is inherent to longitudinal surveys of students whether or not the longitudinal survey is combined with cross-sectional surveys. Calvin Jones's proposal for the creation of testing centers might ameliorate this problem (Jones 1986).

2. Maintaining Acceptable Levels of Nonresponse

A second potential sampling problem with merging a longitudinal and a cross-sectional program has is the possibility that nonresponse will be greater if a cross-sectional survey is embedded in a longitudinal survey than if repeated independent cross-sectional surveys are conducted. This problem can arise because longitudinal surveys run the risk of cumulative nonresponse by students, leading to increasing levels of nonresponse with each successive wave. It is useful to distinguish three types of wave nonresponse, which have been referred to as attrition, re-entry, and late entry (Little and David 1983). A student who drops out of the survey at one wave and does not participate again exhibits attrition nonresponse; reentry occurs when a student who had not participated in one or more previous waves does participate in the later waves; late entry occurs when a base-year nonparticipant does participate in a follow-up.

If most of the nonresponse in longitudinal surveys was attrition nonresponse then we would see nonresponse rates increasing with successive waves. In fact, however, the types of nonresponse are all significantly present. For example, each follow-up in the National Longitudinal Survey of the class of 1972 (NLS-72) and High School and Beyond had lower weighted nonresponse rates than did the base-year surveys, showing that increases in late entry can outweigh increases in attrition. On the other hand, each successive follow-up after the first one suffered larger nonresponse then the preceding. The conclusion

to be drawn is that nonresponse rates in longitudinal surveys do not have to be higher than nonresponse rates for cross-sectional surveys of students in schools.

It is useful to distinguish various causes for student nonparticipation in school-based surveys. In the base-year of a longitudinal survey, and in every year of repeated independent cross-sectional surveys, a student can only participate if, at a minimum, agreement is secured from the relevant chief state school officer, local district superintendent, and school principal. If any of these three educational officials refuses permission, the student will not be able to participate. Of course, the student may also refuse to participate or be absent from a school at all attempted contacts.

These possible sources of nonparticipation affect the base-year of a longitudinal survey and all cross-sectional surveys, but the experience of High School and Beyond suggests that once the education officials permit a base-year longitudinal survey to be conducted, they are also prone to permit later follow-ups to the same survey to be fielded. Since a multiyear longitudinal survey may need permission from fewer principals and district superintendents than will an equivalent number of repeated independent cross-sectional surveys, it may be possible to obtain better school-level participation in a longitudinal survey than in cross-sectional surveys.

3. How Could NAEP Be Made More Longitudinal?

NAEP exemplifies a repeated cross-sectional design. As a primary use of NAEP data is to assess change, both across cohorts and within cohorts, analyses of change would benefit enormously from a more longitudinal design. NAEP could be made more longitudinal in two ways, by keeping schools and school systems in the sample for multiple times or by testing the same students at multiple times. The primary advantage of longitudinal testing of students would be the ability to correlate student growth with school practices and other factors. Such analyses, however, are better performed with NELS surveys, which include more detail on school and home environments. The primary advantages of longitudinal testing in schools and school systems are more precise estimates of change in performance, at the school level, at the district level, at the State level, and nationally.

Longitudinal assessment of schools and districts is well accomplished with a rotation sample design. Given the enormously high intraschool correlation for student ability, I would not be surprised to see at least a doubling of the effective sample size for estimates of two-year change if a rotation sample of schools were adopted. If a rotation sample of districts were used as well then estimates of "cohort" change (e.g., change from grade four in 1986 to grade eight in 1990) would also be vastly improved.

The increase in precision afforded by rotation designs over independent repeated cross-sectional designs will depend on the correlation between a school's aggregate test score on two successive testing occasions. For tests one year apart, Bock (1986) provides evidence that the correlation will be between about .60 and .95, depending on how the test instrument is constructed. For estimates of year to year change, a rotation design will increase the effective sample size by 75 percent if the correlation is .60 and by 700 percent if the correlation is .93; see Appendix C. These potential gains obviously are enormous. While there is some concern (mistaken, in my opinion) that a rotation design would lead to biased test measurements, there can be no question but that the rotation design warrants experimental testing to assay its practicality.

It is less widely known (but true nonetheless) that a rotation design can also improve the precision of simple cross sectional estimates (i.e., not just estimates of change). If the correlation is .60, the increase in effective sample increases by 10 percent and if the correlation is .93 the increase is 36 percent; see Appendix C. These facts further strengthen the case for rotation sampling.

4. How Could a NELS Survey Be Embedded in NAEP?

A variety of ways exist to merge a NELS survey into the NAEP design. For the purposes of this discussion we may assume that NAEP will sample students in grades 4, 8, and 12 every two years. The NAEP sample in any one of these three grades could serve as a base-year sample for a NELS sample. It is also possible to subsample or augment the NELS base-year sample for a NAEP sample.

The primary advantage in merging the two samples in this way will stem from the additional achievement data at the school level that NAEP can provide if a design such as Darell Bock's "duplex design" is used. The duplex design (Bock, 1986) can produce accurate estimates of student ability at the school level if at least 30 students per school are tested. The increased accuracy may be of some value for studying instructional effects; although one should note that NELS studies are poorly suited for assessing the effectiveness of teaching. As students receive instruction from one teacher for a period of a year, a more powerful method is to test the students prior to the instructional period and at the end of this period (e.g., spring-spring testing, or fall-spring testing) and to correlate the achievement gains with teaching method. Such a design is used in the International Educational Assessment (IEA) surveys. Of course, randomized assignment of teaching method would lead to yet more powerful studies of instruction, but even with such an experimental approach, moving towards the IEA design would be valuable. Thus, estimates of school-wide ability can be provided with high accuracy. That can be of value for the study of school effects. Furthermore, use of analytical techniques, empirical Bayes methods and hierarchical modeling in combination with the

duplex design will, I believe, greatly strengthen the pupil-level ability estimates. Several particular mergers will now be considered.

NAEP Grade 12 and NELS Grade 12. The NAEP sample of high school seniors can serve directly as a base-year sample for a NELS survey. Such a merger would be rather straightforward.

I was asked to consider the further possibility of merging the base-year sample with a NPSSAS (National Post-Secondary Student Aid Study) survey. NPSSAS is based on a sample of post-secondary institutions and a sample of students within those institutions. If the NELS base-year sample were essentially identical to the NAEP 12th grade sample, and if the first follow-up survey were conducted at the same time as the NPSSAS sample then it would be possible to have some degree of overlap between NAEP, NELS, and NPSSAS.

The maximum overlap would occur if the selection of postsecondary institutions in NPSSAS were NELS students enrolling in them.

However, as students from a single secondary school will tend to attend a variety of postsecondary institutions, it seems unlikely that a good overlap could be effected. Indeed, the phenomenom of students in one school attending a variety of schools after graduation is problematic enough for grade 6 - grade 10 or grade 8 - grade 10 transitions, and it will be even more severe for post secondary attendance. I do not think that the NPSSAS sample of institutions can be successfully modified to improve the capture of NELS students.

Once the NPSSAS institutions are selected, it is possible to try to maximize the number of NELS students attending those institutions who get selected for NPSSAS. The Keyfitz (1951) procedure for maximizing overlap will apply here provided that the NELS selection probabilities can be determined for all students who attend the selected post-secondary institutions. If the students' NELS base-year selection are known then the selection probabilities can be roughly estimated. To precisely estimate the selection probabilities would require knowledge of the numbers of eligible students in the base-year schools (including numbers in oversampled groups, if any, and whether the student belonged to one of the oversampled groups). Although it might suffice to perform these determinations for a sample of students in the selected postsecondary institutions, the work involved would be considerable.

Thus, to increase the overlap between the NELS and NPSSAS student samples is possible but may not be practical.

NELS:88 Grade 8 and NAEP Grade 12. In 1992 the NELS:88 sample will consist mostly of 12th graders, and a probability sample of 12th graders will be part of the NELS:88 survey in that year (see section 1 and Appendix B for related discussion). The

NELS:88 12th grade sample could be used as the core of a NAEP 12th grade sample but a problem would arise in that the NELS:88 seniors will probably spread out in more schools than if a NAEP 12th grade sample were drawn directly. Thus, the costs of merging the NELS:88 and NAEP 12th grade samples in 1992 might be onerous.

NAEP Grade 12 and NELS Grade 10. Another way to merge the NAEP 12th grade sample with the NELS is to draw the sample of NAEP 12th grade schools two years before the actual testing is to be performed. Then a sample of 10th grade students in those schools could be selected for a base-year NELS sample of 10th graders. The overwhelming majority of the 10th grade sample would attend the same schools two years later, and the sample could then be modified to provide a representative sample of 12th graders. These 12th graders could form the NAEP 12th grade sample. Since the selection probabilities for the schools would be known, the 17-year-old sample could be drawn from the same schools. Thus, this kind of merger seems very feasible.

NAEP Grade 8 and NELS Grade 10. It is also possible to merge a NELS 10th grade base-year sample with a NAEP 8th grade sample. First, a sample of 10th grade schools could be drawn for the NELS base-year sample two years before NELS is to be fielded. Second, the 8th grade schools that "feed" into the selected 10th grade schools could be identified. Third, the 10th grade schools that these 8th grade schools feed into are determined. Once the latter 10th grade schools are identified, the selection probabilities for the selected 8th grade schools may be calculated. The selected 8th grade schools can form the basis for the NAEP 8th grade sample. The sample of 10th grade schools can form the first stage of the NELS 10th grade sample.

Many of the NAEP 8th grade students will, two years later, be enrolled in NELS base-year schools. It would be desirable to use the Keyfitz procedure to maximize the overlap between the NELS students and the NAEP student sample, but the NAEP selection probabilities would need to be known for all 10th grade students in the NELS base-year schools. To determine these probabilities would require identifying the school attended by each student two years earlier and calculating its selection probability. Given the way the 8th grade schools are selected, calculating those selection probabilities appears problematic.

NAEP Grade 8 and NELS Grade 8. Another way to merge the two studies is to let the 8th grade sample of NAEP compose the base-year sample of NELS.

The sampling problems engendered by this approach are no different than those that will be encountered by NELS:88, the main one being that the 8th grade students may spread out to numerous 10th grade schools and make followup costly.

NAEP Grade 4 and NELS Grade 6. One may also merge a NAEP 4th grade sample with a NELS 6th grade sample two years later. First, select a sample of schools containing 4th grade students. Then use the Keyfitz procedure two years later to draw a sample of schools containing 6th grade so as to maximize the overlap with the NAEP 4th grade sample. Use the Keyfitz procedure again at the second stage of sampling to maximize the number of NAEP 4th graders selected into the NELS sample. Note that we would need to determine the NAEP selection probabilities for all 6th grade students in the NELS schools, requiring some data collection for those 6th grade students who attended a different school in grade four. This does not seem problematic, however.

5. Conclusion

Various mergers of NAEP and NELS are possible. The simplest mergers involve merging the base-year of NELS with a NAEP sample at the same grade. It would be rather straight forward to merge a NELS 10th grade (base-year) sample with a NAEP 12th grade sample two years later, or to merge a NAEP 4th grade sample with a NELS 6th grade sample two years later. Other mergers appear problematic.

Critical questions that need to be addressed in designing any merger are, What analyses would be possible with a merged program that are not possible without a merger? What would we learn from the data? The kinds of mergers considered here do not promise many analytical benefits unless the studies are strengthened in other ways as well. For example, testing students twice within a 12 month period would allow measurement of growth, which could be statistically associated with teaching methods and other school processes. If such analyses were possible then the improved testing data afforded by NAEP would have real value for NELS. However, unless we can specify the analytic gains we need from the merger, we might end up with an inferior design.

Appendix A

Following is a brief description of a linking technique, suggested by Martin Frankel, for obtaining a representative sample of 10th grade students in 1990 from a base-year sample of 8th grade students in 1988. The context is NELS:88 but the technique applies to other educational longitudinal surveys as well.

The 8th grade student population will mostly, but not entirely, be enrolled in 10th grade in 1990. To keep the description simple we will assume that all schools enrolling 10th grade students in 1990 enroll some 10th graders who were 8th graders in 1988. First, pick a subsample (or all) of the base-year sample who are enrolled in 10th grade in 1990. Consider any single school enrolling at least one of those students and list all its 10th grade students in a circular list (e.g., alphabetically from A to Z followed by A again). Identify each student in the list who was in the base-year sample and go down the list to the next student. If that student was an 8th grader in 1988, stop; otherwise include that student in the 10th grade sample and go down the list to the next student. If that student was an 8th grader in 1988, stop, and otherwise include that student in the 10th grade sample and go down the list to the next student, etc. Repeat this procedure for all sampled 1988 8th grade students who were in 10th grade in 1990 and for all bschools. To determine sample weights, note that the selection probability for a newly selected 10th grader is equal to that for the sampled 8th grader closest up on the list.

Appendix B

Following are several alternative sampling designs for obtaining not only a representative sample of 10th grade students in 1990 from a base-year sample of 8th grade students, but also a representative sample of schools enrolling 10th graders in 1990. These designs are more complicated than the linking design described in Appendix A. They are presented in the context of NELS:88 but the principles extend to more general kinds of longitudinal studies. The designs will permit accurate inferences to be extended to five major groups or "target populations":

Population A: "1988 8th grade students," i.e., students in the 8th grade in 1988

Population B: Persons in February 1990 who were 1988 8th grade students

Population C: "1988 8th grade schools," i.e., schools in 1988 that contained 8th grade students

Population D: "1990 10th grade students," i.e., students in the 10th grade in 1990

Population E: "1990 10th grade schools," i.e., schools in 1990 that contained 10th grade students

The sample design should possess several important properties.

1. The design should provide a statistically efficient sample of 8th grade students in 1988.

2. The design should allow for cost-effective follow-ups of a subsample (or even all) of the 1988 8th grade students two years later, in 1990.

3. The design should provide a statistically efficient sample of 1988 8th grade schools.

4. The design should provide a statistically efficient sample of 1990 10th grade students.

5. The sample of 1990 10th grade schools should be statistically efficient and also overlap considerably with the 10th grade schools attended in 1990 by the 1988 8th grade students. The overlap will bring about cost reductions in a 1990 survey.

6. The 1990 follow-up sample of 1988 8th grade students should be articulated with the 1990 sample of 10th grade schools to provide a sufficient "density" of the student

sample in schools and districts. For analysis of school-level processes for which no adequate models yet exist (e.g., student transitions from 8th grade to 10th grade, student dropout patterns, grouping/tracking of students), we need to have a sufficient number of observations on students in each school and district in the sample.

7. The sample will yield increased precision for comparisons with the 1980 10th grade class and the 1980 10th grade schools to the extent that the samples of 1990 10th grade schools and of 1980 10th grade schools overlap. Furthermore, an overlap will permit longitudinal analyses of school changes.

These objectives may be achieved with a variety of designs that integrate the 1990 sample of 10th grade schools and students with the 1988 sample of 8th grade schools and students. Four alternative designs will be presented. These alternatives were developed with the collaboration of David Bayless, Michael Brick, James Coleman, Martin Frankel, Morris Hansen, Benjamin King, Benjamin Tepping, and Joseph Waksberg.

Design I

This design yields several samples, which we will label sample A, sample B, and so forth, to correspond to target population A, target population B, etc., listed above. First we give a brief overview of the design and then we describe it in more detail.

First choose a sample of 1988 8th grade schools using probabilities proportional to size within strata (sample C). From these schools, choose a sample of students to yield sample A. Determine which 1988 10th grade schools are "fed" by each sampled 8th grade school (these are the 10th grade schools where sufficient numbers of students from the selected 8th grade schools usually later enroll). Subsample these 1988 10th grade schools. For discussion purposes, we will refer to this subsample as sample S; sample S does not correspond to any target population of interest. Augmenting sample S with 1990 10th grade schools not "fed" by 8th grade schools and screening out non-1990 10th grade schools (if any) yields sample E of 1990 10th grade schools. (These non-1990 10th grade schools would include schools that in 1988 contained 10th grade students but in 1990 did not.)

In 1990 take the sample S and list all the 1990 10th grade students who were in sample A and who also followed one of the identified feeder paths. (These students are not the same as all students in sample A attending 10th grade in 1990 in a school in sample S. Consider, for example, that school X "fed" only to school Y, but Mary Smith in school X attended 10th grade in

school Z, and schools Y and Z were both in sample S. Mary Smith would not be on the list because she did not follow the identified feeder path.) Adding to this list a subsample of sample A who did not follow one of the identified feeder paths yields sample B. This subsample should include all school dropouts, but only a subsample of students who did not follow an identified feeder path. Take sample B, exclude persons who are not 10th grade students, and augment with a subsample of 1,990 10th grade students in sample E schools who were not in-scope 8th grade students in 1988. This yields sample D.

Design I is complicated to describe, but it is really just a well-integrated set of basic component sampling schemes. We next describe the second stage of the sampling in more detail.

The second stage of the sampling begins with an examination of the sampled schools and the determination of which 10th grade schools are "fed" by 8th grade schools. We might characterize a 10th grade school as being "fed" by the 8th grade school if more than some percentage (say, for example, 20 percent) of the 8th grade students are expected to attend that school. Which 10th grade schools are "fed" will be determined in consultation with appropriate authorities, for example, the school district, the Catholic diocese, the private school association, the school principal or headmaster, and so forth.

The 10th grade schools which are "fed" by a sampled 8th grade school will be eligible for selection into sample E, the sample of 1990 10th grade schools. Since the number of these 10th grade schools might be larger than the number desired for sample E, a third stage of sampling to reduce this number may be needed.

The "fed" 10th grade schools can be stratified by the same characteristics used to stratify the 8th grade schools or by other characteristics; they can also be subsampled in such a way as to make their probabilities of selection close to pps. The result will be a probability sample of 1988 10th grade schools that are "fed" by 1988 8th grade schools; earlier we referred to this sample as sample S. (This sample is not useful by itself, but it will be a critical component of Sample E.) To obtain a representative cross-sectional sample of all 1990 10th grade schools, we must augment sample S with a sample of 1990 10th grade schools that are not "fed" by any 1988 8th grade schools.

One way to accomplish this is by sampling primary sampling units (PSUs) which are either school districts or areas such as counties or minor civil divisions. The 8th and 10th grade bschools in the PSUs should be canvassed and those 10th grade schools that are not "fed" by 8th grade schools sampled. It is possible to use the school districts or areas that are implicitly selected in sample C, provided that we form composite districts where unified school districts are not present. An alternative is to draw an independent sample of schools or of areas, such as counties or groups of counties. The resulting sample, sample E,

is a representative (after weighting) sample of all 1990 10th grade schools that can be used for estimating 10th grade school characteristics in 1990.

The student sampling plan uses sample A in two ways to develop samples B and D. Sample B, the follow-up of the 8th grade cohort, is composed of two groups. One group consists of all members of sample A who attend a school in sample E, provided the student followed the identified feeder path. The other group consists of a subsample from those members of sample A who did not follow an identified feeder path into 10th grade. (This subsample could, e.g., be designed to include all of the dropouts, and subsamples of students who are "out-of-grade" or not in a "fed" 10th grade.) These two groups together make up sample B. Diagram B1 clarifies the relationships of the various samples.

Sample D, the 1990 10th grade cohort, also is composed of two groups. As before, one group consists of all members of sample A who attend a school in sample E, provided the student followed the identified feeder path into 10th grade. The other group, however, consists both of 1990 10th grade students who were in 8th grade in 1988 but who did not follow one of the feeder paths and of 1990 10th grade students who were not in-scope 8th grade students in 1988. This last group includes persons who were out of the country in 1988, who were in schools excluded from sample A in 1988, who did not follow the typical grade progression from 1980 to 1990, and so forth. This group will be sampled from the schools in sample E.

Several aspects of this sample design should be noted. The first is that sample selection probabilities can be computed in a straightforward way. This feature is present because the feeding patterns, and hence the sample E schools, are determined only on the basis of the C sample (of schools) and not the sample of students.

Two other aspects of the design depend on plausible assumptions that would need to be investigated. One assumption is that feeder paths exist and can be identified. A second assumption is that we can identify 10th grade students in 1990 who were not in the 8th grade in 1988 as well as 10th grade students in 1990 who did not follow a specified feeder path to their 10th grade school.

Finally, we note that the strong interdependence of samples A through E implies that adjusting the selection probabilities (hence, weights) for one sample has implications for the weights for other samples, and tradeoffs would need to be considered. For example, subsampling of 10th grade schools "fed" by the 8th grade schools (to yield sample S) could be performed to make the schools' selection probabilities close to pps, but this affects the selection probabilities for the follow-up sample (sample B).

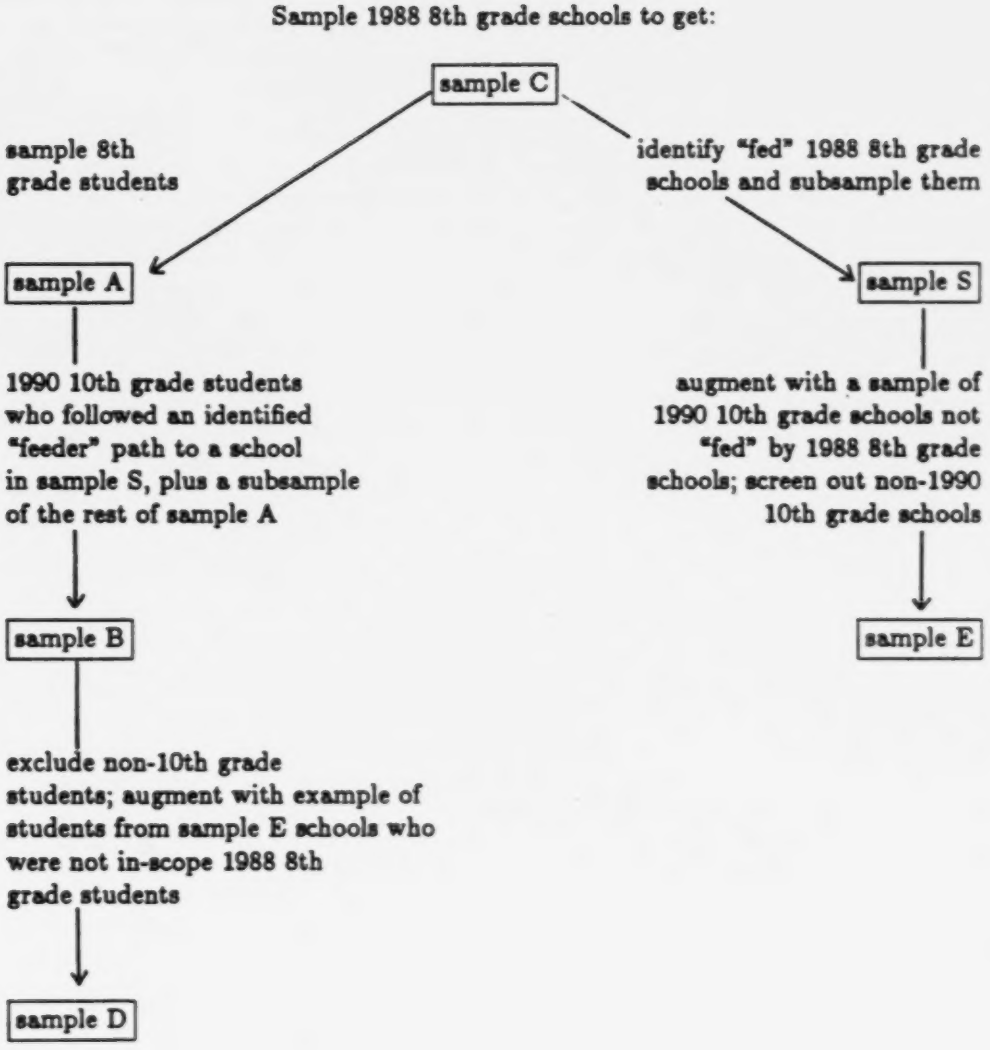

Sample 1988 8th grade schools to get:
sample C
sample 8th grade students
identify "fed" 1988 8th grade schools and subsample them
sample A
sample S
1990 10th grade students who followed an identified "feeder" path to a school in sample S, plus a subsample of the rest of sample A
augment with a sample of 1990 10th grade schools not "fed" by 1988 8th grade schools; screen out non-1990 10th grade schools
sample B
sample E
exclude non-10th grade students; augment with example of students from sample E schools who were not in-scope 1988 8th grade students
sample D

Diagram B.1 Schematic Representation of Design I

Design II

The first stage of sampling is the same as for Design I and it will yield samples A and C. Sample B is obtained by following up in 1990 a subsample of the members of sample A. A person's retention probability (i.e., conditional probability of selection into sample B given that the person was selected into sample A) could be specified in a variety of ways. Equal selection probabilities maximize some measures of information about population B, but their effect will be to spread out sample B among large numbers of 10th grade schools, and hence to greatly increase the cost for 1990 sample. Alternatively, one could try to specify the retention probabilities so as to keep the number of 1990 10th grade schools in the sample small relative to the size of the sample B. One way to do this is to first select the E sample so that it includes large numbers of persons from sample A, and then retain members from sample A who either (i) attended 10th grade in an E sample school, or (ii) did not attend 10th grade in 1990 in a school that was eligible for selection in the E sample. To make this point more clearly, we turn to consider selection of the E sample. Reference to Diagram B.2 will help clarify the different sampling processes.

Each member of the A sample in 1990 will either attend 10th grade in a 10th grade school that is in-scope (i.e., in the population) for the 1990 sampling or will not. Through follow-up and locating activities, we need to determine for each member of the A sample whether he or she is attending 10th grade and, if so, where. Those members not attending 10th grade are subsampled for retention into the B sample. Let us refer to these members as the B1 sample.

Next, the schools in which A sample members are attending 10th grade in 1990 are listed. From this list, schools are selected with probabilities approximately proportional to the number of sample A members who are enrolled in 10th grade there. These selected schools will compose part of the E sample and we will refer to them as the E1 sample.

All (or conceivably a subsample) of the A sample members attending 10th grade in an E1 sample school will be retained in the B sample. Let us refer to these members as the B2 sample. Thus, the B sample consists of the B1 sample and the B2 sample.

To complete sample E, it may be necessary to select a sample of in-scope 1990 10th grade schools that enroll no 10th grade students who were in-scope 8th grade students in 1988. It seems likely that there will be no such schools, but if there are such schools, refer to the sample as the E2 sample. The E sample will consist of the E1 sample schools and the E2 sample schools (if any).

Sample 1988 8th grade schools to get:
sample C
sample 8th grade students
sample A
list schools attended by sample A members as 1990 10th grade students
subsample those not enrolled in 10th grade in 1990
unlabeled sample
sample B1
subsample schools
sample E1
augment with sample of inscope 1990 10th grade schools enrolling no 10th grade students who were inscope 1988 8th grade students
subsample all 10th grade students in sample E1 schools who were in sample A
sample B2
sample E
sample B (B1 ∪ B2)
sample 1990 10th grade students who were not inscope 1988 8th grade students
sample D1
sample D (D1 ∪ B2)

Diagram B.2 Schematic Representation of Design II
∪ denotes union of two samples

To develop sample D we need to sample students from schools in sample E who were not in-scope 8th grade students in 1988. The selection probabilities for these students can be specified in various ways. Call this sample D1. Sample D will consist of sample D1 and sample B1. (It may be desirable to augment sample D with a sample of students attending 10th grade in schools that had low probabilities of selection into sample E1. This augmentation, suitably carried out, would reduce the variation in the weights for sample D. Alternatively, the selection probabilities for the E1 sample of schools could be modified to vary non-linearly, instead of proportionally, as indicated above, with the number of sample A members enrolled in the school.)

This design has advantages as well as disadvantages relative to Design I. The primary advantage is greater control over the sample of 1990 10th grade schools and students (samples B, D, and E). A possible disadvantage is that the selection probabilities for sample E1 schools may be difficult to compute: they require that we ascertain, for each school enrolling a sample A member in 10th grade in 1990, which 8th grade schools each member of the school's 10th grade class attended in 1988.

An alternative to direct calculation of school selection probabilities is to estimate the weights by classifying all 10th grade schools by criteria such as size and location, counting E1 sample schools in each cell in the classification, and computing weights accordingly. Even if we could adequately set up the cross-classification, these weights would only be estimates of the inverses of actual selection probabilities. Thus, this alternative method of calculating selection probabilities also is problematic.

Designs III and IV allow the simplest method for calculating selection probabilities.

Design III

This design selects 8th grade schools (sample C) and 10th grade schools (sample E) essentially at the same time. This is accomplished by specifying the approximate "feeding" patterns for all schools before the sample is drawn. The design offers the possibility of maximizing the overlap between HS&B 1980 10th grade schools and the 1990 10th grade schools in sample E. Such overlap would permit longitudinal analysis of schools and increased precision for comparisons between the sophomore classes of 1980 and 1990.

The first step in this design is development of a list of all in-scope schools in the U.S. that contain an 8th grade and all schools that contain a 10th grade. We will "link" 8th grade schools to 10th grade schools on the basis of prior information about flows of students. In principle, 8th grade schools could

link to more than one 10th grade school. However, for the purpose of explaining the design, we will assume that the schools on the list can be grouped into clusters such that each school belongs to exactly one cluster and such that each cluster contains at least one 8th grade school and one 10th grade school. Ideally, the clusters will contain a few schools, the clusters will contain roughly equal numbers of 8th and 10th grade students, and relatively few students will attend 8th grade in one cluster and 10th grade in another. Each cluster may contain schools with only one type of administrative control; e.g., a cluster may contain only public schools, or only private Catholic schools, and so forth. The present school district boundaries and attendance areas within large metropolitan districts offer good starting points for development of the public school clusters, but we recognize that there may be difficulties in the actual execution of this design.

Once the list is partitioned into these clusters, we can sample clusters with probability proportional to the number of 8th grade students. Alternatively, should inferences about populations D and E be given priority over A through C, we could sample clusters with probability proportional to the number of 10th grade students. Either way, it is possible to use the "Keyfitz" technique of maximizing sample overlap to maximize the extent to which HS&B 1980 10th grade schools are selected for the sample of 1990 10th grade schools (sample E). With this design we stratify clusters in a similar way that we would stratify individual schools. Once a sample of clusters is selected, we can sample 8th grade schools with pps and the same allocation across strata that we discussed for Design I. This sample provides the C sample. From each sampled 8th grade school we would sample 27 students, except for supplemental sample schools from which we would sample 32. This sample provides the A sample. Diagram B.3 clarifies the various sampling processes.

The B sample will consist of a subsample of the A sample. The E sample of 1990 10th grade schools will consist of two parts, an E1 and an E2 sample. The E1 sample will be developed by sampling 10th grade schools with pps from the selected clusters. The E2 sample will be selected from a list of 1990 10th grade schools that were not 10th grade schools in 1988. If the clusters have been designed successfully (to meet the criteria mentioned earlier), then most of the 8th grade students in the A sample will attend 1990 10th grade schools that are in the E (or E1) sample. Thus, the 1990 B and E samples will be efficient ones.

Finally, the D sample of the 1990 sophomore class will consist of two groups. One group will be a sample of students from E sample schools who were not in-scope 8th grade students in 1988; call this sample D1. The other group will be those members of the B sample who attend schools in the E sample; call this sample D2.

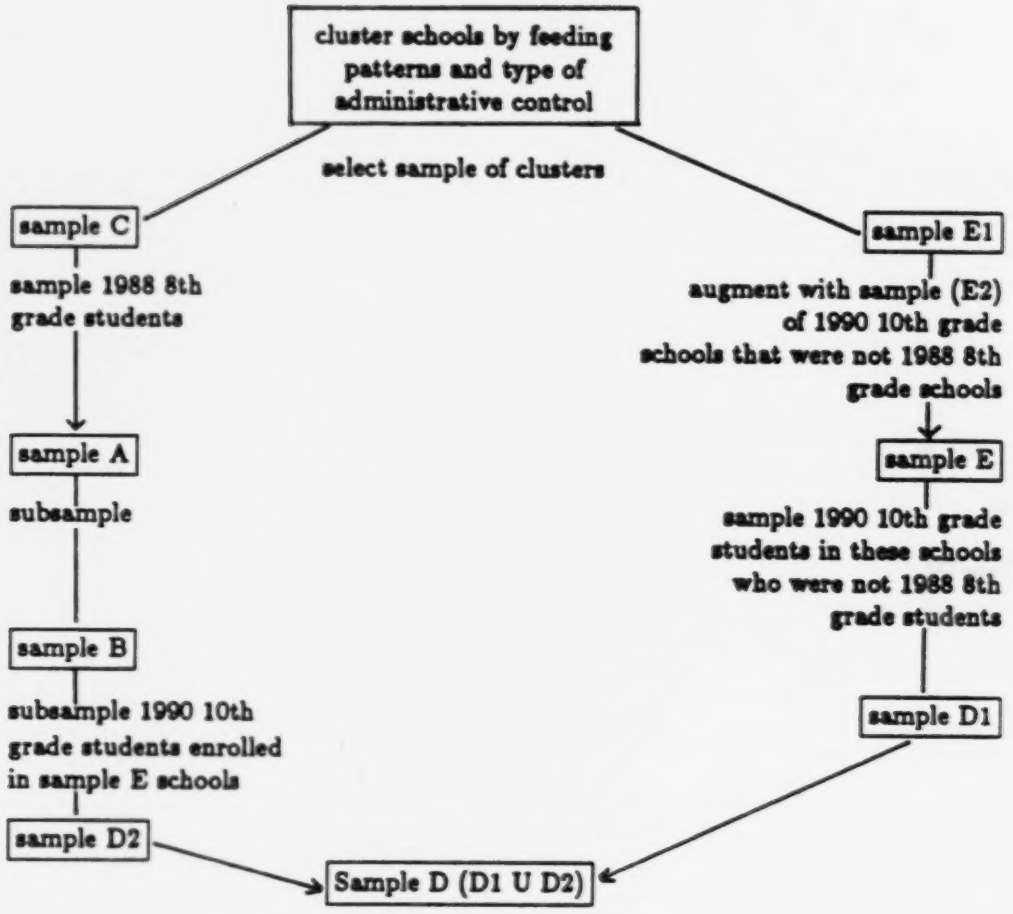

cluster schools by feeding patterns and type of administrative control
select sample of clusters
sample C
sample 1988 8th grade students
sample A
subsample
sample B
subsample 1990 10th grade students enrolled in sample E schools
sample D2
sample E1
augment with sample (E2) of 1990 10th grade schools that were not 1988 8th grade schools
sample E
sample 1990 10th grade students in these schools who were not 1988 8th grade students
sample D1
Sample D (D1 U D2)

Diagram B.3 Schematic Representation of Design III
U denotes union of two samples

The main drawback to this design is the possibile difficulty of determining efficient clusters. Note that if the clusters turn out to be inefficiently developed, then the A sample students will mostly attend 10th grade in schools not in the selected clusters. In this case the E sample just described would be inefficient, but it would still be possible to use Design III to choose the 8th grade schools (sample C) and Design I to choose the 10th grade schools (sample E). With this alternative, "fall back" procedure, if the clusters are unsuccessfully determined then no loss in precision results, while if the clusters are determined efficiently then a substantial overlap with HS&B 10th grade schools can still be achieved.

Design IV

This design is a variant of design III that differs primarily in the clustering and stratification of schools. With this design we need to "redraw" school district boundaries so that they become geographically exhaustive and mutually exclusive. It might be desirable to partition large (in numbers of students) school districts into smaller component "districts". These school "districts", which effectively would be areas, would form the primary sampling units (PSUs). As with Design III, a sample of PSUs would be selected with probability proportional to the number of 8th grade students (pps). The "districts" (PSUs) could be stratified according to various characteristics (see discussion of design I) but not according to type of control, since public and private schools would intermix in the same districts.

A sample of 8th grade public schools would be drawn with 8th grade students from the selected primary sampling units. Then a sample of private schools would be selected with 8th grade students from a subsample of those primary sampling units.

This design has two enormous virtues that arise from selecting private and public schools from the same "districts." First, it will enhance comparisons in the base year between public and private schools because many pairs of public and private schools will come from the same district and hence will be matched on community characteristics. Second, the 1990 school sample can contain many origin (8th grade) and destination (10th grade) schools for the public-private and private-public transitions of students between 8th and 10th grade.

The design shares the other advantages (e.g., 10th grade school overlap with HS&B) and disadvantages (difficulty of determining "districts") of design III.

Appendix C

Increasing Precision by Rotation Sampling

Rotation sampling is a technique to improve the precision of recurring surveys. A technique well-known to sampling experts, rotation sampling has long been used in major government surveys by the U.S. Bureau of the Census, for example in the Current Population Survey (Hanson, 1978). Thus, while rotation sampling would represent an innovation for NAEP, it is not a new invention. The technique is described in standard textbooks on sampling (Hansen, Hurwitz, and Madow, 1953; Kish, 1963; Cochran, 1977). Rotation sampling is still being introduced into some surveys (e.g., the Census Bureau's Retail Trade Survey, as described by Wolter [1979]). The following discussion illustrates how rotation sampling might provide efficiency increases of up to 250 percent or more for estimating year-to-year changes in achievement. Efficiency increases of 25 to 30 percent or more might be attained for estimating achievement levels in any single year.

By "rotation sampling" we mean specifically that schools will enter the sample, be tested on two successive occasions, and then will leave (rotate out) of the sample. Technically, this is known as one-level rotation sampling with 50 percent overlap. Alternative rotation schemes, which would carry their own advantages are certainly possible but will not be considered here. For clarity of exposition the following discussion will simplify somewhat, in that schools are assumed to have equal numbers of students enrolled and sampled, schools are assumed to be selected by simple random sampling, and all variances and covariances are assumed constant. The calculations of relative efficiency can be shown to be unaffected by these assumptions, which greatly facilitate exposition.

The rotation design will now be described (subject to above simplifications). Let $x(\tau, i, \alpha)$ refer to the measured achievement level of school i in year τ, and α denotes whether this is the first or second consecutive appearance of the school in the sample (α may equal 1 or 2 only). Imagine that 2M schools are sampled each year, as follows:

YEAR		
0	1	2
x(0,1,2)		
.		
.		
.		
x(0,M,2)		
x(0,1',1)	x(1,1',2)	
.	.	
.	.	
.	.	
x(0,M',1)	x(1,M',2)	
	x(1,1",1)	x(2,1",2)
		.
	.	.
	.	.
	x(1,M",1)	x(2,M",2)
		x(2,1"',1)
		.
		.
		.
		x(2,M"',1)

Note that schools 1, ..., M appear in sample year 0 only; schools 1' , ... M' appear in the years 1 and 2; schools 1", ..., M" appear in year 2 but not years 0 or 1.

Let τ denote the correlation (or autocorrelation) between the scores for a school on two successive occasions. The great advantage or rotation sampling derives from the autocorrection τ. Calculations based on data from the California Assessment show that, for year to year change, τ could be as small as .6 or as large as .931 (Bock, 1986, Table 1). For estimating change from year 0 to year 1 we may use a so-called composite estimator (Cochran (1977), Hansen (1978), Wolter (1979)).

$$M^{-1}\,(2-r)^{-1}\,[(1-r)\Sigma x(2,i",1) - x(1,i,2) + \Sigma x(2,i',2) - x(1,i',1)]. \quad (1)$$

Similar estimators have been used in the Current Population Surveys since the 1950s. The estimator reduces to the simple average change if $\tau = 0$, that is

$$.5M^{-1}\,[\Sigma x(2,i",1) + \Sigma x(2,i',2) - \Sigma x(1,i',1) - \Sigma x(1,i,2)]. \quad (2)$$

We will refer to estimator (1) as the composite estimator and estimator (2) as the simple estimator.

The relative efficiency of rotation sampling is defined as the number of schools that would need to be sampled if no rotation sampling were used, if we wanted to attain the same precision attainable with rotation sampling based on 100 schools. Thus, relative efficiency is the effective sample size equivalent to 100 schools in the rotation sample design. The relative efficiency depends on the correlation r and the form of the estimator, i.e., composite or simple. For the composite estimator the relative efficiency is given by the formula

$$\text{relative efficiency} = 100 + 50r/(1-r). \qquad (3)$$

For example (see Table C.1, below) if the correlation $r = .85$ and the composite estimator is used then 100 schools in a rotation design give the same precision as 383 schools in a non-rotation design; even if the simple estimator (2) is used, efficiency gains of over 70% can be expected.

TABLE C.1

Relative Efficiency in Estimates of Yearly Change: Various Levels of Correlation r

Correlation r	Current Design (no rotation)	Rotation Design (simple estimator)	Rotation Design (composite estimator)
.60	100	143	175
.70	100	154	217
.80	100	167	300
.85	100	174	383
.90	100	182	550
.95	100	190	1,050

Rotation sampling can also improve the precision of estimates of the level of achievement in a single year. This improvement is based on composite estimator, such as

$$M^{-1}(4-r^2)^{-1}[(2-r^2)\bar{x}(1,\text{I}',1) + 2\bar{x}(1,\text{I}',2) + r\{\bar{x}(0,\text{I},2) - \bar{x}(0,\text{I},1)\}] \qquad (4)$$

for estimating achievement in year 1. Discussion of the rationale behind this estimator may be found in the books by Kish (1963) or Hansen, Hurvitz, and Madow (1953), and its use in the Current Population survey is described in Hanson (1978).

For estimating the level of achievement in a single year the relative efficiency of the composite estimator compared to the simple average achievement in the current year is given by the formula

$$\text{relative efficiency} = 100 + 50r^2/(2-r^2).$$

For example (see Table C.2, below) if the correlation τ equals .85 and the composite estimator is used then 100 schools in a rotation design give the same precision as 129 schools in a non-rotation design.

TABLE C.2

Relative Efficiency In Estimates of Current Achievement: Various Levels of Correlation r

Correlation r	Current Design (no rotation)	Rotation Design with Composite Estimator
.60	100	111
.70	100	116
.80	100	124
.85	100	129
.90	100	134
.95	100	141

The estimators described above will not be used exactly as specified above because the correlation τ will not be known exactly. However, we will be able to specify τ fairly closely, with the effect that our tabulated relative efficiencies will still be approximately correct. For example, if we mistakenly used $\tau = .75$ in forming the composite estimator (4) when in fact τ was truly .85 then the relative efficiency would be 127 rather than 129.

References

Bock, R. D. (1986a) "Designing the National Assessment of Education Progress to Serve a Wider Community of Users". Chicago: NORC.

Bock, R. D. (1986b) "Instrument Design for a Combined NAEP and NELS." Paper presented at a Conference on NAEP and the Longitudinal Studies Program, Center for Education Statistics, Washington, D.C., December 11, 1986. "How to Optimize and Articulate a Longitudinal and a Cross-Sectional Research Program." Paper presented at a Conference on NAEP and the longitudinal Studies Program, Center for Education Statistics, Washington, D.C., December 11, 1986.

Hanson, M. H., Hurwitz, W. N., and Madow, W. G. (1953) Sample Survey methods and Theory. (2 vols.) New York: Wiley.

Hanson, R. H. (1978) The Current Population Survey-Design and Methodology. Technical Paper #40, Washington, D.C.: U.S. Bureau of the Census.

Jones, C. C. (1986) "How to Optimize and Articulate a Longitudinal and a Cross-Sectional Research Program." Paper presented at a Conference on NAEP and the longitudinal Studies Program, Center for Education Statistics, Washington, D.C., December 11, 1986.

Keyfitz, N. (1951) "Sampling with Probability Proportional to Size; Adjustment for Changes in Probabilities," Journal of the American Statistical Association, 52:503-10.

Kish, L. (1965) Survey Sampling. New York, John Wiley & Sons.

Little, R. A. and David, M. H. (1983) "Weighting Adjustment for Nonresponse in Panel Surveys," working paper. Washington D.C., U.S. Bureau of the Census.

Tourangeau, R., McWilliams, H., Jones, C., Frankel, M. F., and O'Brien, F. (1983) High School and Beyond First Follow-Up (1982) Sample Design Report. Chicago: NORC.

Wolter, K. M. (1979) "Composite Estimation in Finite Popula tions," Journal of the American Statistical Association, 74: 604-613.

Alexander, Mary Ellen
National Association of
Secondary School Principals

Amico, Lorraine
National Governors' Association

Ashwick, Helen
Center for Education Statistics

Baron, Joan Boykoff
Connecticut State Department
of Education

Bayless, David
Westat, Inc.

Beaton, Albert
Educational Testing Service

Betka, Sue
U.S. Department of Education

Bock, R. Darrell
University of Chicago

Brown, George H.
Center for Education Statistics

Burkheimer, Graham
Research Triangle Institute

Campbell, Valencia
National Education Association

Carroll, C. Dennis
Center for Education Statistics

Cowan, Charles
Center for Education Statistics

Crowley, Michael
National Science Foundation

Darling-Hammond, Linda
RAND Corporation

Datta, Lois-Ellin
Government Accounting Office

Elliott, Emerson
Center for Education Statistics

Engel, Penelope
Educational Testing Service

Eroe, Melanie
National Computer Systems

Estes, Gary
Northwest Regional Educational Laboratory

Finn, Chester E.
Office of Educational
Research and Improvement

Garduque, Laurie
American Educational Research Association

Ginsburg, Alan L.
U.S. Department of Education

Guhl, Nora
Decision Resources, Inc.

Hall, Ron
Center for Education Statistics

Hansen, Morris
Westat, Inc.

Hedrick, Terry
Abt. Associates

Jones, Calvin C.
National Opinion Research Center

Jordan, Forbis
Congressional Research Service

Kaplin, Gail
National Conference of State Legislators

Karweit, Nancy
Johns Hopkins University

Kaufman, Phillip
Center for Education Statistics

Kolstad, Andrew
Center for Education Statistics

Koretz, Daniel
Congressional Budget Office

Lapointe, Archie
National Assessment of Educational Progress

Lawson, Lawrence
National Computer Systems

McLaughlin, Joan
Government Accounting Office

Manno, Bruno
U.S. Department of Education

Milton, Toby
University of Maryland

Myers, David
Decision Resources, Inc.

O'Neill, John
National Computer Systems

Orr, David
Center for Education Statistics

Owen, Eugene
Center for Education Statistics

Owings, Jeffrey
Center for Education Statistics

Pelavin, Sol
Sol Pelavin Associates

Pendleton, Audrey
Center for Education Statistics

Peterson, Paul
Brookings Institution

Phillips, Gary W.
Center for Education Statistics

Ralph, John
U.S. Department of Education

Rock, Don
Educational Testing Service

Rudner, Lawrence
U.S. Department of Education

Schmidt, William
National Science Foundation

Selden, Ramsay
Council of Chief State School Officers

Song, Tongsoo
Center for Education Statistics

Spencer, Bruce D.
Northwestern University

Stedman, James
Congressional Research Service

Stiglmeier, John
New York Department of Education

Stohl, Adam
American Educational Research Association

Suter Larry
Center for Education Statistics

Sweet, David
Center for Education Statistics

Takai, Ricky
U.S. Department of Education

Treacy, Maureen E.
Center for Education Statistics

Wise, Lauress
American Institutes for Research

Wiley, David
Northwestern University

Wood, Ann
National Computer Systems

Wright, Douglas
Center for Education Statistics

Joan Boykoff Baron, a former high school English teacher, is Director of the Connecticut Assessment of Educational Progress Program. She holds a Ph.D. in measurement and evaluation from the University of Connecticut. She authored or co-authored articles on educational evaluation which appeared in major educational journals such as Educational Measurement: Issues and Practices.

R. Darrell Bock is one of the foremost statisticians in the United States. He received his Ph.D. from the University of Chicago and is a professor of behavioral sciences at that institution. He is a fellow in the American Statistical Association and a past president of the Pyschometric Society. He has written scores of articles on statistics and psychometrics.

Alan L. Ginsburg is the Director of Planning and Evaluation at the U.S. Department of Education. He has participated in the development of many legislative proposals and has published articles on a variety of academic areas. He holds the Ph.D. degree from the University of Michigan. He recently directed the preparation of Secretary Bennett's Schools Without Drugs for which the Department received requests for an unprecedented 900,000 copies.

Calvin C. Jones is a senior survey director at the National Opinion Research Center and is experienced in large survey research. He served as project director of High School and Beyond for about five years and is project director of the National Education Longitudinal Study of 1988 (NELS:88). Mr. Jones is a doctoral candidate at the University of Chicago.

Bruce D. Spencer is Director of the Methodology Research Center at the National Opinion Research Center, an associate professor of education, statistics and policy at Northwestern University, and a sampling statistician at the University of Chicago. He has published extensively on statistical methodological problems. He holds a Ph.D. in statistics from Yale University.

• U. S. GOVERNMENT PRINTING OFFICE: 1987-193-047

Pat Coulter
OERI/IS/ERIC
Mail Stop 1235/Rm. 202A